Cascaded Networks for Thyroid Nodule Diagnosis from Ultrasound Images

Xueda Shen\textsuperscript{1,2}, Xi Ouyang\textsuperscript{1(✉)}, Tianjiao Liu\textsuperscript{3}, and Dinggang Shen\textsuperscript{1(✉)}

\textsuperscript{1} Shanghai United Imaging Intelligence Co., Ltd., Shanghai, China
demo.ouyang@gmail.com, Dinggang.Shen@gmail.com
\textsuperscript{2} Department of Mathematics, University of Illinois Urbana-Champaign, Champaign, IL, USA
\textsuperscript{3} Department of Electronic Engineering, Tsinghua University, Beijing, China

Abstract. Computer-aided diagnostics (CAD) based on deep learning methods have grown to be the most concerned method in recent years due to its safety, efficiency and economy. CAD’s function varies from providing second opinion to doctors to establishing a baseline upon which further diagnostics can be conducted [3]. In this paper, we cross-compare different approaches to classify thyroid nodules and finally propose a method that can exploit interaction between segmentation and classification task. In our method, detection and segmentation results are combined to produce class-discriminative clues for boosting classification performance. Our method is applied to TN-SCUI 2020, a MICCAI 2020 challenge and achieved third place in classification task. In this paper, we provide exhaustive empirical evidence to demonstrate the applicability and efficacy of our method.

Keywords: Thyroid nodule · Ultrasound images · Detection · Segmentation · Classification

1 Introduction

Computer-aided diagnostics (CAD), especially in thyroid nodule classification task have a long history. Following huge boost in image classification performance, people started fine-tuning existing networks to classify thyroid nodules [5]. Such fine-tuning, even though exhibits decent performance on certain data sets, could not achieve a universally optimal performance since the fine-tuned networks could be weak in extrapolation. On the other hand, directly fine-tuning networks for classification often misses out entirely on segmentation, which still strains doctors in diagnostics. To address this, there has been an emergence of interest in detection and segmentation methods [4]. On the other hand, there are also questions with regards to whether segmentation should be included since it requires significant computational resources and is relatively easy for radiologists to segment thyroid nodule from image.
There could be 3 tasks involved for this problem: 1) detection, 2) segmentation, and 3) classification. Ma et al. [1] developed a hybrid model for automatic nodule detection and segmentation. Specifically, it first employs a deep neural network to learn probability maps around ground-truth area. Then, all the probabilities maps are split by the splitting method. Another CNN segments the image from these maps. However, these methods assume a Bernoulli distribution for generating probability maps, which could be questionable. There has been an abundance of literature focusing on thyroid classification. However, most of these methods focus on a somewhat coarse approach by only fitting a pre-trained network to a data set. In the paper by Li et al. [5], they fine-tuned ResNet-50 on a data set featuring different cohorts of ultrasonic thyroid images. Even though they have achieved great performance, this could be partially attributed to the size of their training set with $N = 42952$, while each validation set is only around 1000 images. The disparity in size makes such approach virtually useless in clinical applications. For the classification part, Song et al. [2] created a Multitask Cascade Convolution Neural Network for integrating segmentation and classification. The network features a two-step design. VGG-16 is used as backbone to extract feature maps and recognize nodule coarsely. After this, a spatial pyramid based recognition network finely segments and classifies the nodule. This work integrates segmentation and classification tasks but fails to utilize information generated in segmentation process to aid classification.

In our paper, we answer this question with detailed empirical evidence. Finally, we propose a cascaded network that exploits inherent cues from detection, and segmentation tasks to achieve the final classification prediction that has high sensitivity and specificity, alleviating the workload of doctors.

2 Method

Detection and Segmentation Architecture. Our method aims to complete detection, segmentation and classification. We employed a picture level ensemble strategy by ensembling on masks generated by Mask Scoring R-CNN [9] and CentreNet [11] + Deep Snake [12] combination. In the first branch we employed Mask Scoring R-CNN as a candidate for mask prediction. This network is able to generate high quality masks. However, since the mask generation requires a high threshold, the network produces empty mask for some hard cases in our experiments. To address this issue, we added a two-step segmentation mechanism featuring CenterNet for detection and Deep Snake for segmentation, compensating for lack of detection in Deep Snake. CentreNet [11] is a one stage method for object detection. The network enriches information by centre pooling and cascade corner pooling which mitigates the issue of corner points not capturing image information, thereby increasing detection performance. On the other hand, Deep Snake [12] is an instance segmentation method that is based on circular convolution and contour deformation. It has fast segmentation speed and gives competitive performance. However, due to contour deformation, the mask generated by Deep Snake can sometimes be too smooth. To mitigate inherent
drawbacks of both networks, we added a MLP module which consisted of three conv layers for mask selection as shown in Fig. 1.

**Fig. 1.** Demonstration of ensembled segmentation work flow. Two masks are generated by Mask Scoring R-CNN and two-staged segmentation method consisting of CenterNet and Deep Snake.

**Fig. 2.** Two-step attention network. CBAM module is responsible for telling the network where to focus. CAM is responsible for guiding feature maps.

**Classification Architecture.** Attention mechanism could be used to boost classification performance [13,20]. Class activation map (CAM), proposed by Zhou et al. [13], is able to produce class discriminatory information. The work in its original form, however, allows us to understand the network more but is unable to be directly used to increase the performance since the attention map generated is not integrated in the training process. The method used by Ouyang et al. [14] integrates CAM in an online manner that further boosts the performance of classification network. Convolution Block Attention Module, CBAM, proposed by [17] establishes a method to produce channel features, essentially telling the network “where” to focus. Our network,
whose architecture is demonstrated in Fig. 2, utilizes both module. We call this method two-step attention mechanism. After feature maps are generated, the first stage is to produce channel features by CBAM. The online CAM module generates attention map under the guidance of mask, which the second step of our attention mechanism.

We use ResNet-34 as backbone of our network for feature extraction. The inputs of our network are the heatmaps generated by CenterNet of respective category, the original image and a channel featuring aspect ratio which is an important indicator when diagnosing malignancy. After the feature maps are generated, they are forwarded into the channel attention module proposed in CBAM. Figure 3 denotes channel attention architecture. Letting $\mathbf{F}$ denote the feature map generated by our ResNet-34 backbone, of dimension $\mathbb{R}^{c \times w \times h}$, the channel-wise attention module will produce a channel feature map denoted by $\mathbf{M}_c$ of dimension $\mathbb{R}^{c \times 1 \times 1}$. Formally, it is generated by:

$$\mathbf{M}_c = \sigma(\mathbf{F}_{max}^c + \mathbf{F}_{avg}^c),$$

where $\sigma$ denotes the sigmoid activation. After the generation of this channel-wise attention, we apply this by

$$\mathbf{f} = \mathbf{M}_c \odot \mathbf{F},$$

where $\odot$ denotes element-wise multiplication and $\mathbf{f}$ denotes the feature map after applying channel attention.

The feature map and weights of the last fully connected layer undergoes $1 \times 1$ convolution to generate the attention map. Formally, let $\mathbf{f}$ denote feature maps and $\mathbf{w}$ be the weight matrix of the fully connected layer. Attention map $A$ is given by:

$$A = ReLU(conv(\mathbf{f}, \mathbf{w})).$$

The attention map will therefore be of the same shape with any channel of the feature map. Attention map is then upsampled to the original input size and undergoes color normalization. We then perform softmasking with sigmoid function:

$$T(A) = \frac{1}{1 + exp(-\alpha(A - B))},$$

where $T(A)$ is the attention map generated by this online attention module. Furthermore, this online module designed a combined loss so that we can calibrate both attention map and our classification results, i.e.,

$$Loss = L_{classification} + \lambda L_{dice}$$

where

$$L_{classification} = BCEloss$$

We use Dice loss to maximize the overlap between attention map and input mask. The classification loss is set to be Binary entropy loss. $\lambda$ provides a lever-aging effect between the two tasks; and since classification is the main task, we set $\lambda = 0.4$. 

Fig. 3. Channel feature architecture. The input feature are mapped to max pooling and avg pooling separately, then passing through a MLP network.

It should be noted that in training process, the weights of $1 \times 1$ convolution layer is an identity map from those of the fully connected layer. The weights of the convolution layer is only updated by $L_{\text{classification}}$ since $L_{\text{dice}}$ skips the GAP layer in back propagation.

This online, learnable, channel focus CAM module is able to improve the performance of our network and explicitly states area of interest learned by the network. This explainable factor makes the network more interpretable in its decision making process and would further increase the credibility of the network.

3 Experiment

3.1 Data Set and Augmentation Techniques

TN-SCUI 2020 data set features 3644 ultrasound images of Thyroid gland, of which 2003 are malignant and 1641 are benign. The data set is provided by courtesy of Shanghai Ruijin Hospital. The dataset is then partitioned into training and validation in a 7:3 ratio. To further increase the robustness of our method, we employ a variety of data augmentation methods. Specifically, we randomly rotate the image and apply small degrees of affine transformation to mimic the positions and hardware variances in image acquisition process. Furthermore, we increase the diversity of our data by adjusting brightness, contrast and Gaussian noise. Finally, we train the network on a five-fold, cross validation and cast a majority voting on the testing set featuring 910 images.

3.2 Ablation Study on Ensembled Segmentation

We compare the results of our ensembled segmentation method with those of other models utilized in this paper. The results are shown in Table 1. In particular, we evaluate all of our networks on the testing data set given by the organizers and achieved a 0.3% increase in Mean IoU.
Table 1. Segmentation result of Ensemble UNet. DLA34 stands for Deep Layer Aggregation Model with 34 layer, and FPN stands for Feature Pyramid Network. DLA34 denotes Deep Layer Aggregation Model with 34 layer.

| Model                                                   | Mean IoU (%) |
|---------------------------------------------------------|--------------|
| Deep Snake (CenterNet Detector, DLA34 Backbone)         | 76.71        |
| Mask Scoring R-CNN (ResNet50 + FPN Backbone)            | 79.28        |
| Ensembled segmentation (ours)                           | 79.58        |

Table 2. Cross comparison of multiple classification methods. SVM classifier denotes SVM classification on HOG, SIFT and Gabor features. TS-ResNet34 denotes our two-step attention mechanism with ResNet34 as backbone.

| Method            | Accuracy (%) |
|-------------------|--------------|
| ResNet            | 75.45        |
| VGG               | 73.82        |
| SVM classifier    | 72.11        |
| Mask R-CNN        | 77.12        |
| ResNest50         | 77.69        |
| CenterNet         | 77.92        |
| TS-ResNet34       | **81.01**    |

3.3 Ablation Study on Classifier

Table 2 presents classification performance of multiple classification networks. Table 3 presents ablation study on the modules featured in our network and Fig. 4 presents the attention map produced by the network with varying modules applied in the network. We established the superiority of our classifier in the following two regard. First, we demonstrate our method has superior performance to established method in this field. On the other hand, we provide empirical evidence suggesting the necessity and edge of our two-step attention module. Furthermore, we have conducted a brief explainability assessment of our network, ensuring our method provides interpretable decision making process to better assist clinical diagnostics.

Combining Table 3 and Fig. 4 gives us a better understanding of our network. Observing images from (2, 1) to (2, 6), we are able to see that without mask guidance, the network was unable to properly guide most of its attention on to the nodule. This situation is best represented by images (2, 4) and (2, 5) as we can that see most of the attention is placed on the perimeter of the image rather than the actual nodule itself. Differences between (2, x) and (3, x), $x \in 1, .., 6$, demonstrate drastic improvements in terms of placement of attention generated by the network. This provides further evidence for inherent correlation between segmentation and classification. Numerically, this is the difference in classification metrics reflected in Table 3, between Serial # b and c, which shows
Table 3. Ablation study of our two-staged attention network. TA stands for two-step, CBAM stands for the usage of the channel attention module, and CAM stands for the usage of CAM attention module. Heat map represents the usage of the class-discriminative detection heat map generated by CenterNet. Ratio represents the addition channel of input consisting of height and width ratio. mIoU represents the mean intersection of union between attention map generated by CAM and segmentation mask.

| Serial # | Method     | CBAM | CAM | Heat map | Ratio | ACC  | F1   | mIoU  |
|----------|------------|------|-----|----------|-------|------|------|-------|
| a        | ResNet34   |      |     |          |       | 0.7341 | 0.7477 | 0.0364 |
| b        | TA-ResNet34| ✓    |     |          |       | 0.7240 | 0.7823 | 0.0171 |
| c        | TA-ResNet34| ✓    | ✓   |          |       | 0.7978 | 0.8267 | 0.5016 |
| d        | TA-ResNet34| ✓    | ✓   | ✓        |       | 0.7896 | 0.8188 | 0.0591 |
| e        | TA-ResNet34| ✓    | ✓   |          |       | 0.7814 | 0.8020 | 0.5990 |
| f        | TA-ResNet34| ✓    | ✓   | ✓        | ✓     | 0.8019 | 0.8343 | 0.5869 |

Fig. 4. Instances of the attention map generated by our two-step attention mechanism. Left three are benign cases and the remaining are malignant. First row represents original ultrasound images of the thyroid nodule. Second row represents the attention maps produced by the network without mask guidance. Third row represents the attention map generated with mask guidance without heat map and width height ratio. Fourth row represents the attention maps generated with mask guidance, as well as, with the heat map and the height width ratio as additional inputs. We denote the leftmost image on the first row by (1, 1) and rightmost image at the same row by (1, 6). Also, denote the rightmost image in the fourth row by (4, 6).
that the refinement of the attention maps leads to improvements in classification performance. Differences between (3, x) and (4, x), x ∈ 1, ..., 6 represents the difference in attention maps with additional inputs. Even though there is a slight drop in mIoU, the attention regions are more closely fitted to the nodule area, and the rate of change in attention intensity is more continuous. Furthermore, additional inputs are able to diminish opportunities of wrongfully identifying nodules. Looking at differences between (3, 3) and (4, 3), the secondary nodule’s attention values are mitigated, which can also be observed between (3, 5) and (4, 5). Such mitigation reduces the risk of misdiagnosis.

3.4 Comparison with Detection-Based Classification

To further illustrate the superiority of the proposed method, we conduct another experiment for this task. We first detect the thyroid nodule with Mask Scoring R-CNN [9] and crop the image according to the proposed bounding box. We then fine tune a network on the cropped images. Specifically, the final classification results depend on the patch images from the detection part.

For detection, we use Mask Scoring R-CNN for proposing the target bounding box [9], which is an improved version of Mask R-CNN [8]. In our method, Mask Scoring R-CNN is not concerned with classification of the nodule, i.e., giving only one category, called Thyroid Nodule. Therefore, after several epochs, s_{cls} comes close to 1, allowing the network to give much of its attention to fine-grained segmentation. Predicted mask is forwarded to Mask IoU head.

For classification, the images are segmented accordingly to the bounding box proposed by Mask Scoring R-CNN. We employed VGG, ResNet, ResNest and Gabor features for comparing classification results. The comparison of classification algorithms are presented by Table 4. The best result is achieved by ResNest, a variation of ResNet employing split attention module [19]. The split attention module produces attention along the channel axis to better highlight useful information for image classification.

Table 4. Comparison of multiple classification methods on the validation set. For neural networks, we adopt weights that are pretrained on the ImageNet. The pretrained networks are then trained on the train set for 30 epochs with an initial learning rate of 0.0001 and undergoing a decay of factor 0.2 every 10 epochs.

| Method      | Accuracy (%) |
|-------------|--------------|
| ResNet      | 75.45        |
| VGG         | 73.82        |
| SVM classifier | 72.11     |
| ResNest50   | 77.69        |

The above results provide empirical evidence that even though split attention module of ResNest is able to boost classification performance, its performance is
still much lower than that of our method (i.e., 81.01% in Table 2). This phenomenon demonstrates that, using the localization cues from detection and the segmentation task is suitable for exploring the performance of this task, while roughly cropping the nodule regions may lead to too much misguidance for the final classification.

4 Conclusion

In this paper, we explored inherent connections between segmentation and classification, and designed a two-step attention network to utilize segmentation results for achieving better classification results. Our method achieved the third place in classification at TN-SCUI2020 challenge. Furthermore, our method provides explainable learning by explicitly producing attention maps generated by the network, which we hope would aid doctors in clinical diagnostic process.

References

1. Ma, J., Wu, F., Jiang, T., Zhu, J., Kong, D.: Cascade convolutional neural networks for automatic detection of thyroid nodules in ultrasound images. Med. Phys. 44(5), 1678–1691 (2017). https://doi.org/10.1002/mp.12134
2. Song, W., et al.: Multitask cascade convolution neural networks for automatic thyroid nodule detection and recognition. IEEE J. Biomed. Health Inform. 23(3), 1215–1224 (2019). https://doi.org/10.1109/JBHI.2018.2852718
3. Castellino, R.A.: Computer aided detection (CAD): an overview. Cancer Imaging 5(1), 17–19 (2005). https://doi.org/10.1102/1470-7330.2005.0018. The official publication of the International Cancer Imaging Society
4. Chi, J., Walia, E., Babyn, P., Wang, J., Groot, G., Eramian, M.: Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network. J. Digit. Imaging 30(4), 477–486 (2017). https://doi.org/10.1007/s10278-017-9997-y
5. Li, X., et al.: Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. Lancet Oncol. 20(2), 193–201 (2019). https://doi.org/10.1016/s1470-2045(18)30762-9
6. Prochazka, A., Gulati, S., Holinka, S., Smutek, D.: Classification of thyroid nodules in ultrasound images using direction-independent features extracted by two-threshold binary decomposition. Technol. Cancer Res. Treat. 18 (2019). https://doi.org/10.1177/1533033819830748
7. Guo, D., et al.: Organ at risk segmentation for head and neck cancer using stratified learning and neural architecture search. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 4222–4231 (2020). https://doi.org/10.1109/CVPR42600.2020.00428
8. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: 2017 IEEE International Conference on Computer Vision (ICCV), Venice, pp. 2980–2988 (2017). https://doi.org/10.1109/ICCV.2017.322
9. Huang, Z., Huang, L., Gong, Y., Huang, C., Wang, X.: Mask scoring R-CNN. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, pp. 6402–6411 (2019). https://doi.org/10.1109/CVPR.2019.00657
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90

11. Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., Tian, Q.: CenterNet: keypoint triplets for object detection (2019). (cite arxiv:1904.08189 Comment: 10 pages (including 2 pages of References), 7 figures, 5 tables)

12. Peng, S., Jiang, W., Pi, H., Li, X., Bao, H., Zhou, X.: Deep snake for real-time instance segmentation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 8530–8539 (2020). https://doi.org/10.1109/CVPR42600.2020.00856

13. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, pp. 2921–2929 (2016). https://doi.org/10.1109/CVPR.2016.319

14. Ouyang, X., et al.: Dual-sampling attention network for diagnosis of COVID-19 from community acquired pneumonia. IEEE Trans. Med. Imaging 39(8), 2595–2605 (2020). https://doi.org/10.1109/TMI.2020.2995508

15. Mayo Clinic: Thyroid nodules (2020)

16. Mayo Clinic: Needle biopsy (2020)

17. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: convolutional block attention module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 3–19. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_1

18. Liu, S., Deng, W.: Very deep convolutional neural network based image classification using small training sample size. In: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), Kuala Lumpur, pp. 730–734 (2015). https://doi.org/10.1109/ACPR.2015.7486599

19. Zhang, H., et al.: ResNeSt: split-attention networks. arXiv:2004.08955

20. Fu, J., et al.: Dual attention network for scene segmentation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, pp. 3141–3149 (2019). https://doi.org/10.1109/CVPR.2019.00326