Coarse to Fine Ensemble Network for Thyroid Nodule Segmentation

Zhe Tang and Jianqiang Ma

Alibaba, Damo Academy, Hangzhou, China
tangzhe0220@icloud.com, mjq93@qq.com

Abstract. A thyroid nodule is defined as a small lump of tissue (either solid or cystic - filled with fluid), usually more than one-quarter of an inch in diameter, that may protrude from the neck’s surface or may form in the thyroid gland itself. The nodule can be either benign (non-cancerous) or malignant (cancerous). By the age of 45, up to half of the normal people have thyroid nodules that can be seen on an ultrasound. Fortunately, about 95% of thyroid nodules are benign. Recently, artificial intelligence becomes more and more popular in medical image processing. It helps the doctor in some scenarios; one is the thyroid nodule analysis in the ultrasound image. In this competition, we design a robust coarse to fine network that reaches a high-performance nodule segmentation result. Our method could handle some problematic cases, especially with no clear boundary of the thyroid nodule.

Keywords: Thyroid nodule · Image segmentation

1 Introduction

Image segmentation is a fundamental technique in medical image processing. It is widely used to segment targets to assist doctors in making clinical decisions. For the thyroid nodule, doctors usually need to diagnosis according to the imaging manifestations features, such as the nodule location, volume size, edge smoothness, etc. Artificial intelligence could help doctors with these diagnoses so that to make a better decision for BI-RADS validation.

The quality of data strongly influences ultrasound image segmentation. Artifacts such as speckle, shadows, missing boundary, and noise usually exist, leading to nodule segmentation difficulty. In this competition, we present a novel method “Coarse to fine segmentation network”, which reaches a 0.8194 IOU performance at the final leaderboard 3rd.

The rest of this paper is organized as follows. Section ‘Method’ describes the idea of this innovation, the architecture of the proposed network, and usage of the dataset. Section ‘Experiments and Results’ describes the dataset and evaluation results, section ‘Conclusion’ summarizes this proposed method.
2 Method

In our method, we present an ensemble image segmentation network. There are four unique algorithms. Two basic segmentation network including deeplabv3+ [1] and U2Net [3] shown as Seg. Network A and Seg. Network B in Fig. 1, while network A based on deeplabv3+ uses two backbones, xception and resnet101 for the ensemble. Another two networks C and D, are described in section ‘Architecture of Coarse to Fine Network’. A two stages of image segmentation networks are used to produce coarse segmentation and fine segmentation. Feature extraction backbone is based on xception in network C and D. The basic workflow is shown in Fig. 1.

![Fig. 1. Overview of workflow](image-url)

2.1 Architecture of Coarse to Fine Network

Due to the difficulty of ultrasound image segmentation, especially the around nodule boundary, only an individual segmentation network can not always get a highly accurate result. Thus, we design a coarse to fine structure, which is a two-stage segmentation algorithm. The coarse to fine network consists of two main parts, coarse segmentation network, and fine segmentation network. The coarse and fine segmentation networks are based on Deeplabv3+ architecture. The coarse network outputs an initial predict mask, which is concatenated with the original image and set as the fine network’s input. These two networks are optimized individually.
Fig. 2. Overview of coarse to fine segmentation network

As shown in Fig. 2, there are two Deeplabv3+ networks. In the first network, the origin ultrasound image with data augmentation is set as input, and output a segment predict mask, then we concatenate with image and set as the input of the fine network.

We use five cross-validations on the first coarse segmentation network, inference segment mask of each case, some of them may have good results, but some are bad. We compare with ground truth. A threshold of 0.75 was used to compare with IOU between predict coarse mask and ground truth; when IOU $\geq 0.75$, we set these masks as the input of fine segmentation network, while the masks with IOU $< 0.75$ is replaced by ground truth with applying the random elastic transformation.

When training on a fine segmentation network, an original image and a coarse mask are concatenated as double-channel input, producing a fine predict mask. The coarse to fine workflow is shown in Fig. 3

As shown in Fig. 1, there are two coarse to fine networks C and D. The difference between them is, network C uses the binary mask as the input of fine segmentation network, while network D uses Gaussian smoothed mask as the input of fine segmentation network. The target of using a Gaussian smoothed mask is to reduce the impact of the wrong boundary, but due to our experiments,
in some cases, binary mask input is better. In contrast, some others with a Gaussian smoothed mask are better, so we use both for the ensemble step.

We tried another coarse to fine workflow. When coarse mask is obtained, a patch will be extracted around the predicted mask, but we have not got a good IOU result in Leaderboard A, so we didn’t use this in the final testing for Leaderboard B.

2.2 Loss Function

We use IOU loss as loss function, the initial loss rate is $10^{-3}$, and we also use loss scheduler function ‘ReduceLROnPlateau’ with 0.6 factor in 10 patience.

We have tried some other loss functions such as BCE loss, boundary loss, dice loss, etc. However, the validation results look very similar.

2.3 Data Preprocess and Augmentation

Data preprocess steps including, image resize to 512 * 512 resolution, intensity normalize to 0–1 range with the float data type.

Data augmentation is significant, especially on the ultrasound image. We test the algorithm based on Deeplabv3+ baseline with and without data augmentation. The IOU metric improves about 9% with augmentation. We use random rotation, random flip, random elastic transformation, and random scale or shift on intensity in our design. Also, random Gaussian noise is applied to the image.
2.4 Inference Test Time Augmentation

Test Time Augmentation (TTA) method is used in our inference step, including original, Rotation 90, Rotation 180, Rotation 270, and their flip images, with a total of 8 predict results for average. TTA improves performance obviously in our experiments, about 2–3%.

2.5 Ensemble

All ensemble step here we used is averaging the predict probability maps. We have tried binary masks but, the result is similar.

3 Experiments and Results

3.1 Dataset

There are 3644 images in training set, during algorithm development stage, 10% cases used for validation and 10% cases used for testing, the other 80% used for training. Table 1 experiments are all based on the training data set. According to these experiments, the IOU metrics result on self-define test set are a little lower than evaluation IOU result with official test data set.

When leadership opened, we separate all training data into 5 cross validation used for coarse network step, inference all predict masks with 5 cross validation models. For other networks, all training data are used in training step. There are 910 images in testing set, all of them are used for testing part.

3.2 Result Evaluation

We tried more than 30 experiments with different baseline, backbone, or parameters during the algorithm development stage. We picked some typical results shown in Table 1.

According to the experiments, Deeplabv3+ has better results than UNet [4] series, PSPNet [5], Hourglass [2], etc. In the table, data aug. means data augmentation including RandRotate90, RandFlip, RandScaleIntensity, RandShiftIntensity, RandGaussianNoise, RandAffine and Rand2DElastic. All of these operations have half probability execution. Because the task is binary segmentation, sigmoid is used here as an activate function.

In post-processing, erode and dilate are also used to remove some outliers.

Some other parameters we set here: batch size = 32 or 64, normalize range = [0,1], image resize to 512 * 512 as input of model.

About the final result, we use 4 individual algorithm branches before ensemble, and their results on leaderboard A are shown in Table 2.
Table 1. Experiment results on self-define test set (365 images)

| Network                                      | Data process             | Loss  | IOU  |
|----------------------------------------------|--------------------------|-------|------|
| UNet                                         | original data            | iou   | 0.6495 |
| UNet                                         | data aug.                | iou   | 0.7417 |
| UNet++                                        | data aug.                | iou   | 0.6799 |
| U2Net                                        | data aug.                | iou   | 0.7817 |
| shape-attention-unet (densenet)              | data aug.                | iou+contour | 0.7584 |
| PSPNet+(resnet101)                           | data aug.                | iou   | 0.7796 |
| Hourglass                                    | data aug.                | iou   | 0.7505 |
| Deeplabv3+(resnet101)                        | data aug.                | iou   | 0.7725 |
| Deeplabv3+(drn)                               | data aug.                | iou   | 0.7734 |
| Deeplabv3+(xception)                         | data aug.                | iou   | 0.7861 |
| Deeplabv3+(xception)                         | data aug.                | bce   | 0.7820 |
| Deeplabv3+(xception)                         | data aug. + heatmap      | iou+bce | 0.7855 |
| Deeplabv3+(xception)                         | data aug. dual channel(img+coarse mask) | iou   | 0.7732 |
| Deeplabv3+(xception)                         | data aug. + TTA          | bce   | 0.7989 |

Table 2. Experiment results on Leaderboard A

| Network                                                  | IOU  |
|----------------------------------------------------------|------|
| Deeplabv3+(xception + resnet101 ensemble) + TTA         | 0.8220 |
| U2Net + TTA                                             | 0.8112 |
| Coarse to Fine Net + TTA                                | 0.8079 |
| Gaussian Coarse to Fine Net + TTA                       | 0.8175 |

4 Conclusion

Our method proposes a coarse to fine two-stage network, which we believe can get good results in ultrasound image segmentation. Also, ensembles are used in the inference step. These improve the final results very much. Data processing is another essential part of the algorithm. Augmentation in training steps with random spatial transform, intensity transform, and TTA are efficient ways to improve results. We tried some methods focusing on nodule boundaries such as boundary loss and attention operations during the competition, which we think should be helpful. However, the result is not better. We think this can be continued to improve in the future.

References

1. Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 833–851. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_49
2. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. CoRR abs/1603.06937 (2016). http://arxiv.org/abs/1603.06937
3. Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O., Jagersand, M.: U2-net: going deeper with nested U-structure for salient object detection. Pattern Recogn. 106, 107404 (2020)

4. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28

5. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890 (2017)