Automatic lung segmentation in low-dose CT image with contrastive attention module

Yang, Changxing, Tian, Haihong, Xiang, Dehui, Shi, Fei, Zhu, Weifang, et al.

Changxing Yang, Haihong Tian, Dehui Xiang, Fei Shi, Weifang Zhu, Xinjian Chen, "Automatic lung segmentation in low-dose CT image with contrastive attention module," Proc. SPIE 11313, Medical Imaging 2020: Image Processing, 1131333 (10 March 2020); doi: 10.1117/12.2548806

SPIE.

Event: SPIE Medical Imaging, 2020, Houston, Texas, United States

Automatic Lung Segmentation in Low-Dose CT Image with Contrastive Attention Module

Changxing Yang[a], Haihong Tian[a], Dehui Xiang[a]*, Weifang Zhu[a], Fei Shi[a], Xinjian Chen[a]
[a]School of Electronic and Information Engineering, Soochow University, Suzhou, Jiangsu Province, 215006,China

ABSTRACT

Automatic lung segmentation with severe pathology plays a significant role in the clinical application, which can save physicians' efforts to annotate lung anatomy. Since the lung has fuzzy boundary in low-dose computed tomography (CT) images, and the tracheas and other tissues generally have the similar gray value as the lung, it is a challenging task to accurately segment lung. How to extract key features and remove background features is a core problem for lung segmentation. This paper introduces a novel approach for automatic segmentation of lungs in low-dose CT images. First, we propose a contrastive attention module, which generates a pair of foreground and background attention maps to guide feature learning of lung and background separately. Second, a triplet loss is used on three feature vectors from different regions to pull the features from the full image and the lung region close whereas pushing the features from background away. Our method was validated on a clinical data set of 78 CT scans using the four-fold cross validation strategy. Experimental results showed that our method achieved more accurate segmentation results than that of state-of-the-art approaches.

Keywords: automatic lung segmentation, contrastive attention map, triplet loss

1. INTRODUCTION

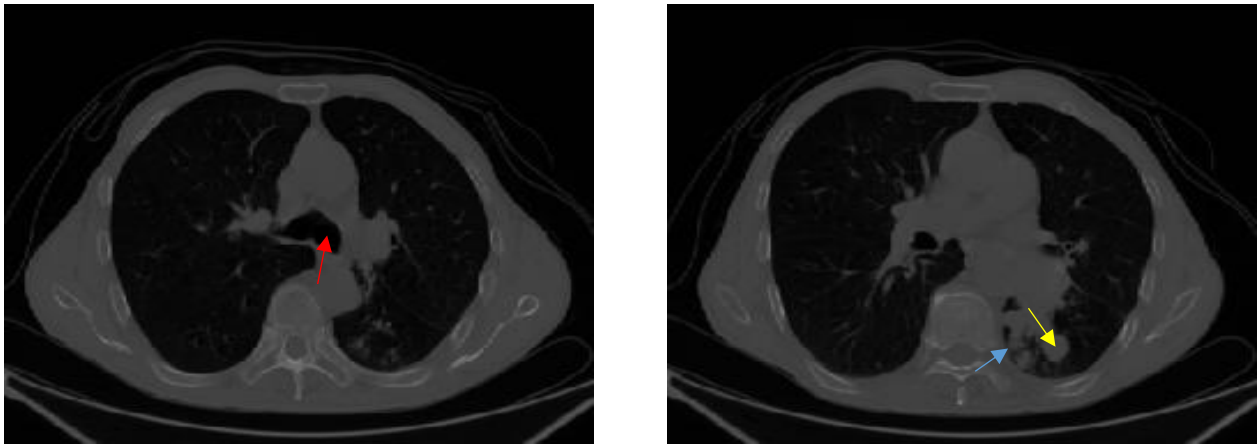
Lung cancer causes high death rates in both men and women [1], and lung segmentation can help to improve the accuracy of lung tumor segmentation and lung nodule detection [2]. As shown in Fig.1, it is still a challenging task to segment pathological lung. First, the contrast between the foreground and background is low due to low dose radiation and the boundaries between tumors and other organs are often unclear. In low-dose CT images, motion artifacts produced by breathing are often evident. Second, large tumors often result in dramatic changes in lung structures and shapes. Third, the anatomy of the lungs varies a lot from different individuals.

Many methods have been proposed for automatic lung segmentation in CT images in the past. Some of them are based on threshold algorithm. These threshold-based segmentation methods performed well in CT images with normal lungs, but poorly in those with large tumors or artifacts. To address these challenges, Sun et al. [3] proposed a novel robust active shape model (ASM) approach to successfully segment the large tumors ignored by traditional threshold methods. Zhang et al. [4] used a deformable segmentation model to segment lung via local sparse shape composition (SSC) on a learned dictionary. Local SSC model was proposed to describe the shape in a segment-to-segment manner which can preserve local details for segmentation task.

In addition, deep learning has been demonstrated to achieve excellent performance in various challenging tasks, such as classification, segmentation and detection. The encoder-decoder model has been shown to be one of the most efficient network architectures for segmentation tasks. UNet [5] is the most common network architecture adopted in medical image segmentation. The encoder gradually reduces the spatial dimension of feature maps and capture the long-range information while the decoder recovers object details and spatial dimension. Skip connections are employed to help decoder layers output a more accurate segmentation result by fusing features from encoder layers. LaLonde and Bagci [6] proposed a convolutional-deconvolutional capsule network (SegCaps) which contains much fewer parameters without decreasing the accuracy for object segmentation.

*Corresponding author: Dehui Xiang, E-mail: xiangdehui@suda.edu.cn

Medical Imaging 2020: Image Processing, edited by Ivana Išgum, Bennett A. Landman,
Proc. of SPIE Vol. 11313, 1131333 · © 2020 SPIE · CCC
code: 1605-7422/20/$21 · doi: 10.1117/12.2548806



(a) (b)

Fig.1. Illustration of the challenges in pathological lung segmentation. (a) Red arrow denotes the organ with the large similarity to the lung. (b) Blue arrow represents the fuzzy boundary and yellow arrow indicates the large tumor.

However, most deep learning methods, such as UNet and SegCaps, directly learn features from the whole image, which contains not only the foreground features, but also the background features. It is effective to focus on the most pertinent piece of information rather than using all available information, so removing the background features during training is helpful for improving the segmentation performance.

In addition, we usually set the number of the filters at each layer to be large in order to capture as many features as possible, so attention mechanisms are often added to CNN for weighting the features. Attention mechanism has achieved great success in various fields including natural language processing and computer vision, such as machine translation [7], image segmentation [8] and object detection [9]. It is effective to carry out a channel attention across channels or a spatial attention across locations. Inspired by attention mechanism, we introduce a contrastive attention module, which generates a pair of foreground and background attention maps to guide feature learning of lung and background separately. We explore to use the pair of attention maps to enhance the foreground features and reduce the background features via a triplet loss. Instead of design an independent module to learn attention without supervision, we simply use the coarse segmentation result from a pre-trained UNet as a mask to guide an attention map to learn key features.

Our contributions are summarized as follows:

•A contrastive attention module is proposed to learn foreground and background features separately.

•We further use a triplet loss to pull the features from the full image and the lung region close whereas pushing the features from background away.

•Our network can achieve more accurate segmentation results than that of state-of-the-art approaches.

2. METHODS

The overview of our proposed method is shown in Fig.2. There are three branches, including the full image, foreground and background. The top branch learns features from the full image. The middle branch learns the lung region features with a foreground attention map under the guide of the coarse result obtained by a pre-trained UNet. On the contrary, the bottom branch learns the background features with a background attention map. The pair of attention maps are generated by a contrastive attention module in the middle, Then, we use a triplet loss on three feature vectors learnt from fully-connected layers of three branches in order to make the features from the lung region and background far away from each other. The final segmentation result is obtained from the foreground branch.

2.1 Data Augmentation

78 3D low-dose CT images are cut into 2D slices for training in our experiments. Considering that the number of slices with lung organs in a whole 3D image is limited, it is necessary to generate a large number of slices to tune the massive network parameters. Thus, we horizontally flip, vertically flip and randomly rotate the slices to make data augmentation during training.

2.2 Network Architecture

Fig.2 shows the network structure of our proposed method. For a given 2D lung image, the first two convolution-blocks of Resnet-34[10] without down-sampling operation produces several feature maps with a size of 390×290×128 whose height and width are the same size as the input image. The feature maps can generate a pair of attention maps by our contrastive attention module. Then the foreground-branch and background-branch are multiplied by the two attention maps to implement spatial attention. All the three branches compute a 128-dimention feature vector using a fully-connected layer after the encoder architecture of UNet, representing the features learnt from full image, foreground and background respectively. The complete UNet contains the encoder-decoder architecture in the foreground branch gets the final segmentation result. The details of the contrastive attention module and loss function are shown in section 2.3 and 2.4.

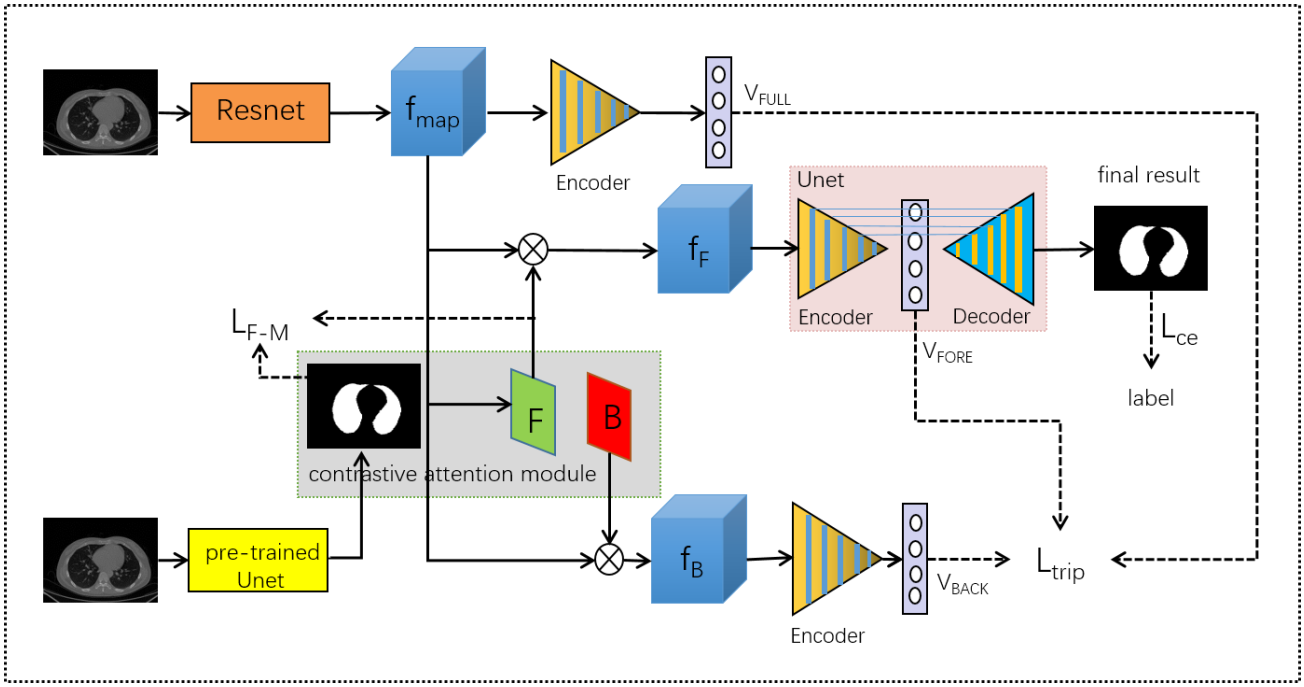


Fig.2. Framework of proposed method for lung segmentation

2.3 Contrastive Attention Module

Spatial attention module is used to produce a weighting map to implement spatial-wise attention across each location in feature maps. In this way, the network could focus on the exactly spatial regions that contribute more for training the model. As illustrated in Fig.2, given an input 2D lung image, the feature maps after Resnet can be denoted as f_{map}. Taking f_{map} as inputs, our contrastive attention module produces a foreground attention map by two convolution operations. The foreground attention map can be denoted as

$$F = \mathrm{S}(\mathrm{W} * f_{map} + \mathrm{b}) \tag{1}$$

where $\mathrm{S}(\mathrm{x}) = 1/(1 + \exp(-\mathrm{x}))$ is the sigmoid function. W and b are the convolutional filter weights and bias. We then generate an inverse map B named background attention map. To ensure that F and B can constitute a contrastive pair, each location (i, j) in F and B should meet the constraint:

$$F(i,j) + B(i,j) = 1 \tag{2}$$

The dimensions of the two attention maps F and B are both 390×290×1. Afterwards, we use this pair of contrastive attention maps as spatial attention for lung and background feature learning:

$$f_F = f_{map} \otimes F \tag{3}$$

$$f_B = f_{map} \otimes B \tag{4}$$

where $\otimes$ means the pixel multiplication. The foreground attention map is expected to concentrate on the lung region. Therefore, the segmentation result from the pre-trained UNet is used as a lung mask to guide the foreground attention map, which guarantees that it certainly learns lung features. The Mean Squared Error (MSE) loss between the foreground attention map and corresponding lung segmentation mask can be defined as:

$$L_{F-M} = \sum_{i=1}^{W} \sum_{j=1}^{H} \| F(i,j) - M(i,j) \|_2^2 \quad (5)$$

where M is the lung mask obtained by a pre-trained UNet, which has the same size of the foreground attention map. W and H are the width and height of the attention map. Therefore, the contrastive attention module can generate contrastive features related to the foreground and background separately.

2.4 Loss Function

We further introduce a triplet loss to enhance the foreground features and reduce the background features. Three 128-dimensional feature vectors, which can be denoted as v_{FULL}, v_{FORE} and v_{BACK} are the representations of the full image, foreground and background. The triplet loss can be defined as

$$L_{trip} = \| v_{FULL} - v_{FORE} \|_2^2 + max\{(m - \| v_{FULL} - v_{BACK} \|_2^2), 0\} \quad (6)$$

where m is a margin parameter which is set to 10 in the experiments. There are two main terms in Equation (6): the first term pulls the foreground features close to the full image and the second term pushes background features far away from the full image. With the minimization of this loss, the features from the full image and the lung region will get close to each other whereas the features from background will be away. In other words, the features from the lung region and background will not intersect each other, which can make the network certainly learn lung features. With $\mathrm{p}(x_i)$ representing the probability prediction of a pixel x_i after the softmax function in the last output layer, the crossentropy loss can be formulated as:

$$L_{ce} = -\sum_{x_i \in \mathcal{X}} y_i \log \mathrm{p}(x_i) \quad (7)$$

where $\mathcal{X}$ represents the output map and y_i is the target class label of the pixel $x_i \in \mathcal{X}$. Taking the MSE loss into consideration, the total loss can be denoted as

$$L_{total} = L_{ce} + \alpha \cdot L_{F-M} + \beta \cdot L_{trip} \quad (8)$$

where α is the balance weight of L_{F-M} and β is the balance weight of L_{trip}, which are respectively set to 0.1 and 0.01 in our experiments.

3. RESULTS

3.1 Datasets

Our segmentation approach was evaluated in 78 3D low-dose CT images obtained from different patients with lung tumors. These images are acquired by a GE Discovery ST16 PET-CT scanner from the top of the skull to the upper part of the femur. CT scanning parameters were 120KV voltage, 150 mA current, and 3.75mm thickness. The pathological lungs were manually labeled as ground truth by clinical experts. All the original CT image size is 512×512×299, we cut them to 390×290×130 and the voxel size is 0.98mm×0.98mm×3.75mm.

3.2 Data Preprocessing and Data Augmentation

In our experiments, we randomly split the 78 CT images into 4 parts, which contains images from 20, 20, 20 and 18 subjects, for the four-fold cross validation. All 3D CT images are cut into 2D slices for training. However, the number of slices with lung organs is limited, we need to do data augmentation on slices which have lungs by flipping horizontally, flipping vertically and rotating randomly.

3.3 Implementation Details

For data augmentation, we implement horizontal flip, vertical flip and rotation via Opencv at each epoch during training. The corresponding scales are {True, True, 30}. Our model was trained on a workstation equipped with one NVIDIA Tesla K40m GPU with 12G memory for 20 epochs. We used Adam optimization with an initial learning rate of 10^{-4} and the batch size was set to 4 in our experiments.

3.4 Segmentation Results

To quantitatively assess the performance of our proposed method, we compared the segmentation results with the ground truth according to the following four metrics: dice similarity coefficient (DSC), true positive fraction (TPF), false positive fraction (FPF) and precision. The DSC calculates the similarity between the segmentation results and ground truth, it is defined as

$$\mathrm{DSC} = \frac{2TP}{FP+2TP+FN} \tag{9}$$

Where, TP is the number of true positives, TF is the number of true negatives, FP is the number of false positives and FN is the number of false negatives. TPF, FPF and precision metrics are computed as:

$$\mathrm{TPF} = \frac{TP}{TP+FN} \tag{10}$$

$$\mathrm{FPF} = \frac{FP}{FP+TF} \tag{11}$$

$$\mathrm{Precision} = \frac{TP}{FP+TP} \tag{12}$$

We compared the segmentation results to different methods as shown in Table 1. In order to evaluate the segmentation results fairly, ASM [3] and SSC [4] methods were trained and tested using the leave-one-out strategy while deep learning methods, including UNet, SegCaps and our method used the four-fold cross validation strategy. Some segmentation results of these methods are shown in Fig.3.

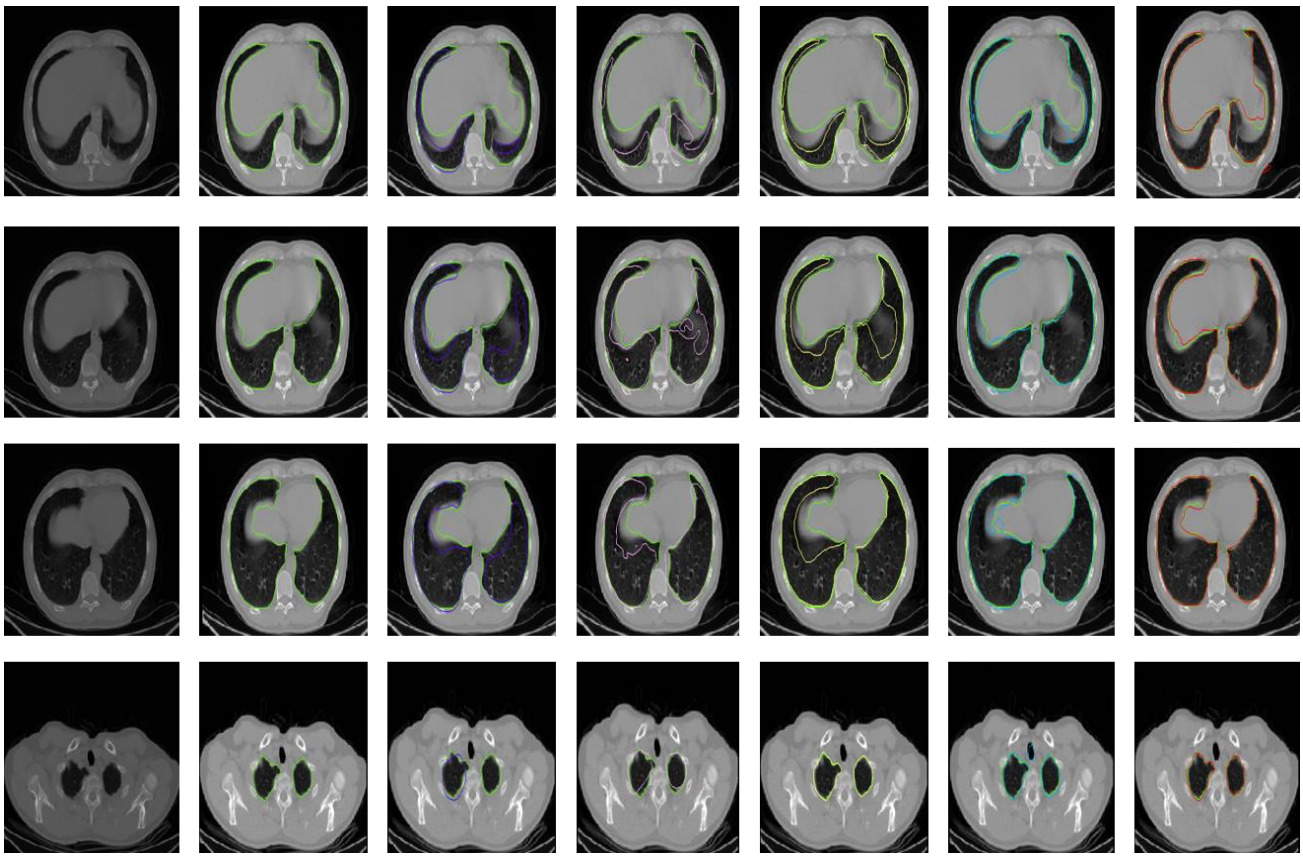


Fig. 3. Segmentation results of different methods. The first column is the original image; the second column is the ground truth; the next columns are the results of SSC (sparse shape composition), ASM (active shape model), UNet and Segcaps; the last column is our segmentation results.

Table 1. Comparison of the quantitative segmentation results for different methods (mean±standard deviation).

Methods	DSC (%)	TPF (%)	FPF (%)	Precision (%)
ASM[3]	94.43±1.55	95.66±2.68	0.23±0.11	93.63±1.87
SSC[4]	94.62±1.37	93.28±2.26	0.16±0.04	89.72±0.74
UNet [5]	95.16±2.16	93.10±2.89	**0.09±0.13**	95.83±1.39
SegCaps[6]	93.81±1.79	**97.05±2.30**	0.42±0.23	90.85±0.77
Our method	**96.62±1.51**	96.33±1.60	0.15±0.09	**96.56±1.32**

4. CONCLUSIONS

In this paper, a new approach for automatic lung segmentation is introduced. First, we propose a contrastive attention module to learn features from the lung region and background separately. Second, a triplet loss is used to pull the features from the full image and the lung region close whereas pushing the features from background away. It is able to enhance the foreground features and reduce the background features. Our method achieves competitive segmentation results compared to state-of-the-art approaches.

5. ACKNOWLEDGEMENTS

This work has been supported in part by the National Key R&D Program of China under Grant 2018YFA0701700, and in part by the National Natural Science Foundation of China (NSFC) under Grant 61971298.

6. REFERENCE

[1] J. Ferlay, I. Soerjomataram, M. Ervik, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. Parkin, D. Forman, and F. Bray, "Globocan 2012 v1. 0," Cancer incidence and mortality worldwide: IARC CancerBase, no. 11, 2013.

[2] S. G. Armato and W. F. Sensakovic, "Automated lung segmentation for thoracic CT impact on computer-aided diagnosis," Acad. Radiol., vol. 11, no. 9, pp. 1011–1021, 2004.

[3] S. Sun, C. Bauer, and R. Beichel, "Automated 3-D segmentation of lungs with lung cancer in CT data using a novel robust active shape model approach," IEEE Trans. Med. Imag., vol. 31, no. 2, pp. 449–460, Feb. 2012.

[4] S. Zhang, Y. Zhan, and D. N. Metaxas, "Deformable segmentation via sparse representation and dictionary learning," Med. Image Anal., vol. 16, no. 7, pp. 1385–1396, 2012.

[5] O. Ronneberger, P. Fischer, and T. Brox, U-net: Convolutional networks for biomedical image segmentation, in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent, 2015, pp. 234–241.

[6] Lalonde R, Bagci U. Capsules for Object Segmentation. 2018

[7] 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, pages 5998–6008, 2017.

[8] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. In CVPR, 2016.

[9] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In CVPR, 2017.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In CVPR, 2016.