Journal of Image Processing Theory and Applications (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/jipta.2023.060105
ISSN 2560-6239 Vol. 6 Num. 1

CT Image Segmentation Network of COVID-19 Based on Multi-scale Attention and Probability Preserving Pooling

Zhiwen Yang[1,2], Mingju Chen[1,2,*]

[1]*School of Automation and Information Engineering, Sichuan University of Science and Engineering, Zigong, 643000, China*
[2]*Artificial Intelligence Key Laboratory of Sichuan Province, Yibin, 644000, China*
[*]*Corresponding author*

Keywords: COVID-19, deep learning, image segmentation, multiscale attention, class probability preserving pooling

Abstract: In order to effectively improve the accuracy of U-Net segmentation of COVID-19 CT images, a novel COVID-19 CT image segmentation model based on multi-scale attention and class probability Preserving pooling auxiliary classifier was proposed. In view of the shortcomings of U-Net in using single-size standard two-dimensional convolution to extract image feature information, a multi-scale attention module is adopted to enhance the performance of U-Net in extracting multi-scale information by using three groups of attention modules mixed with different sizes of deep strip convolution and spatial domain channel domain. Aiming at the disadvantage of information loss when using maximum pooling in the U-Net lower sampling layer, two-dimensional convolution is used to reduce the size of the feature map, and an auxiliary classifier is used to calculate the coarse segmentation semantic probability map of each layer of the encoder and the loss is calculated by using the GT tag map of the class probability Preserving pooling to optimize the classification performance of the network. The network model is tested for image segmentation in the standard COVID-19 CT image dataset, and the segmentation accuracy for the infected area is 87.22%, the recall rate is 84.55%, and the intersection/merger ratio is 75.98%.

1. Introduction

The gold standard for detection of COVID-19 is to use RT-PCR technology, which needs to wait for the results and cannot be quickly screened. On February 4, 2020, the National Health Commission of China issued the Diagnosis and Treatment Plan for Pneumonia Infected by novel coronavirus (the fifth version for trial), which listed CT technology as an important means to detect the characteristics of COVID-19.

Using image segmentation technology, we can extract the infection area of the focus in the CT image of COVID-19 patients, and realize automatic and rapid detection. The characteristics of COVID-19 CT images are that the lesion area is vague, the edge shape is complex, and the lesion area is similar to the background gray intensity. Traditional image segmentation algorithms such as

threshold based image segmentation, edge detection based image segmentation, and wavelet transform based image segmentation algorithm have poor segmentation effect on COVID-19 CT images.

With the improvement of computing ability of computer hardware, deep learning, with its complex network combination and parameter quantity, can fit the network model of huge image data and extract the deep features of the image, surpassing the traditional methods in image segmentation. Long, et al. proposed the full convolution FCN network to solve the problem of too large parameters of traditional convolution neural network[1], and Shelhamer, et al. proposed the U-Net network[2], which has the structure of encoder and encoder, and uses the jump connection to recover the information loss problem of up-sampling and down-sampling. Towaki et al. proposed Gated-SCNN[3], which controls the boundary of the feature graph in the network, generates the semantic segmentation boundary through the gradient and feature of the image, and uses a pixel-based loss function to optimize, making the network more accurate in the boundary shape segmentation. Fanet al. proposed inf-net[4], which uses semi-supervised learning to learn unlabeled samples to solve the problem of less labeling of medical images. Zhu et al. proposed STLNet, which proposed a new operator to quantify and count the low-level texture information in the image features[5]. On this basis, the texture enhancement module was used to enhance the texture details, and the pyramid texture feature extraction module was used to mine the texture information from multi-scale.

The above network models are all aimed at improving a certain disadvantage in the context of deep learning, but fail to improve a network from multiple aspects, resulting in frequent replacement of the backbone network when applying the network model, without a stable and long-term use of the backbone model. In this paper, we will use U-Net, a network commonly used in medical segmentation, as the backbone network, and improve the shortcomings of U-Net to improve the segmentation ability of the model.

2. Segmentation Network Based on Multi-scale Attention and Probability Preserving Pooling

2.1. Overall Network Structure

CT image has fuzzy boundary and complex shape, which requires strong feature extraction ability of network. In this paper, U-Net is used as the basic partition network, and the structure is coder-decoder structure. The encoder is composed of four groups of image feature extraction modules and down-sampling. The image feature extraction module contains two multi-scale attention modules. The multi-scale attention module is used to make up for the lack of multi-scale information extraction in U-Net. Three sets of deep strip convolutions of different sizes are used to replace the standard two-dimensional convolution to improve the network multi-scale information capture ability, and the CBAM (convolutional block attention module) attention module[6] is added to improve the modeling ability of network context information. For the problem of information loss in the down-sampling of U-Net, the maximum pooling is used, and the two-dimensional convolution is used to conduct the down-sampling of the feature map, Then input the down-sampling feature map into the auxiliary classifier and use the GT real map label of the class probability reservation pooling layer to calculate the KL divergence as the auxiliary loss function to optimize the network. The overall structure of the network is shown in Figure 1.

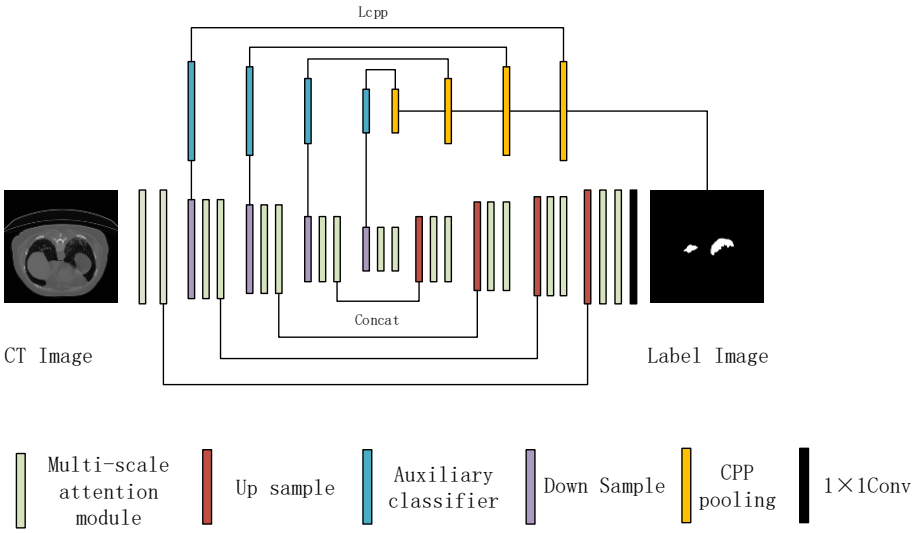

Figure 1: Network structure diagram.

2.2. Multi-scale Attention Module

Multi-scale feature capture is very important for network to improve classification performance, such as Google-Net[7] or HR-Net[8]. Referring to the MSCA[9](multi-scale convolutional attention) module structure, our paper designs a multi-scale attention module to improve the ability of U-Net to extract multi-scale features and global context modeling. The module structure is shown in the figure 2, which is composed of multi-scale feature extraction, attention calculation and residual branch.

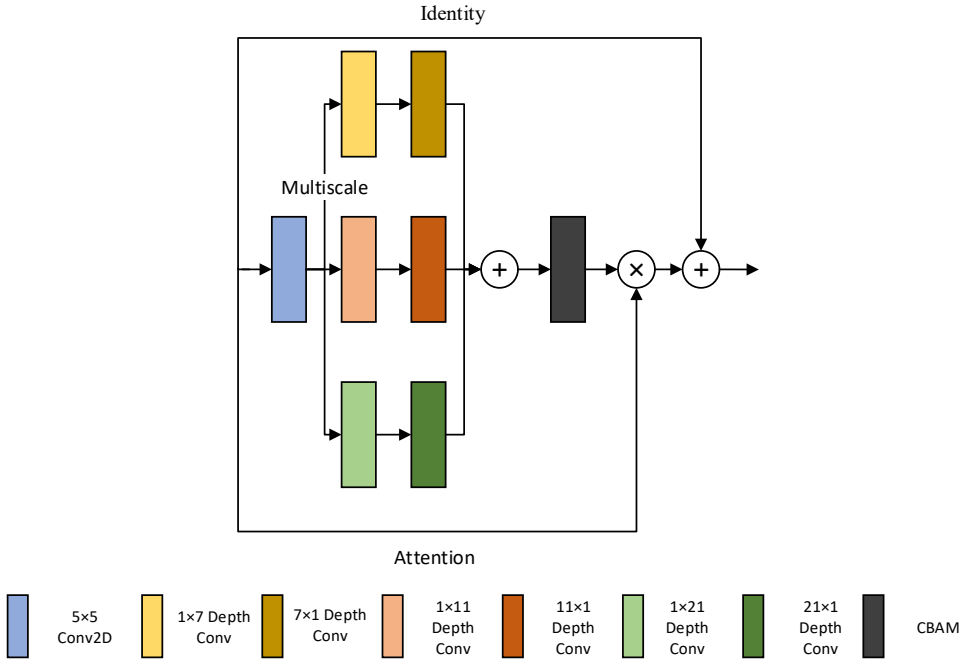

Figure 2: Multi-scale attention mechanism module.

For the input feature map, three sets of depth strip convolutions of different sizes are used to extract multi-scale features. Compared with the standard two-dimensional convolution, the depth strip convolution[10] is better than the standard two-dimensional convolution in parameter quantity, and the feature extraction ability is not significantly different from the two-dimensional convolution, and is

conducive to the extraction of small strip features. Multi-scale feature map uses bitwise addition method for feature fusion and input into CBAM attention module to calculate the attention weight map, and multiply the weight map with bitwise. Finally, the residual structure is used to solve the network degradation problem. In this paper, we use the mixed CBAM of spatial domain and channel domain to make up for the shortage that the MSCA weight calculation only focuses on the channel dimension.

CBAM structure is composed of channel attention module and spatial attention module in series. For the feature map, the attention weight map in the channel domain will be calculated first, and then the attention weight map in the space domain. The channel attention module performs squeezing operation through two parallel operations of average pooling and maximum pooling, and then interacts with channel information through a shared perceptron. The attention weights in the parallel structure are added bit by bit, and finally the channel attention weight is generated through sigmoid function. The weight calculation formula is shown in Formula 1:

$$M_c(F)=\sigma(MLP(Avg(F))+MLP(Max(F))) \quad (1)$$

The spatial attention module is a parallel structure with maximum pooling and average pooling according to the channel dimension. Both operations are to compress the information in the channel domain. The feature map generated by the parallel structure is spliced according to the channel dimension. To increase the receptive field, use 7 × the convolution kernel of size 7 is used to model the spatial information, and the channel dimension is adjusted to 1. Finally, the sigmoid function is used to generate the spatial weight map. The structure diagram of CBAM is shown in Figure 3:

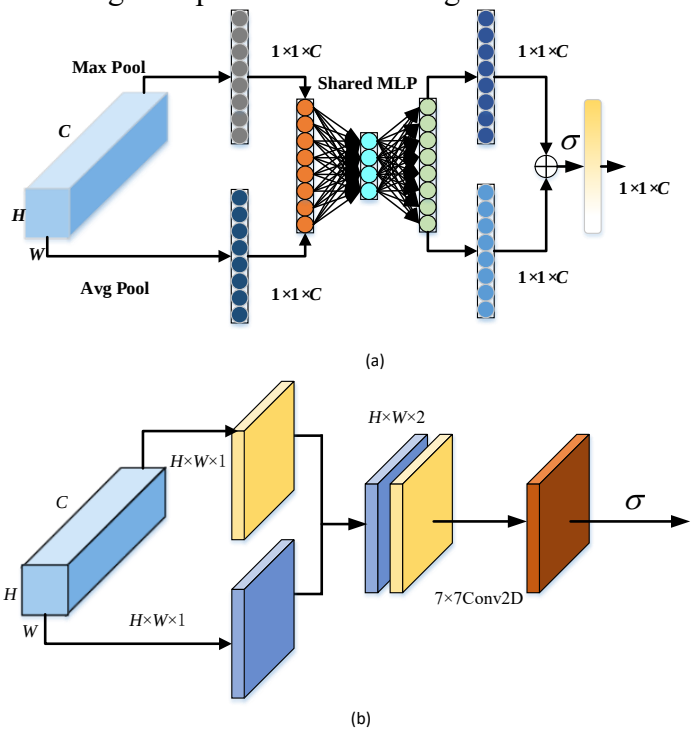

Figure 3: CBAM module structure, (a) is the channel attention module, and (b) is the spatial attention module.

2.3. Auxiliary Classifier Based on Class Probability Preserving Pooling

The maximum pooling layer in U-Net will lose the information in the characteristic graph. Therefore, the maximum pooling down sampling of U-Net is changed to convolution operation, and

the auxiliary classifier is used to optimize the network performance. The use of auxiliary classifiers can optimize the learning process of the network, such as PSPNet[11]. When the auxiliary classifier calculates the auxiliary loss, it usually needs to down-sample the GT tag graph. The common sampling methods are maximum pooling and average pooling. This paper uses the CPP (class probability preserving pooling) structure[12] and optimizes the loss of the GT graph and the split graph. The auxiliary classifier calculates the probability diagram of the rough segmentation diagram, which is composed of 3 × 3-size convolution, BN, Relu, 1 × 1 size convolution, and finally use softmax to calculate the probability diagram. The probability information of each class in the classification task can be retained by using CPP pooling, and the structure is shown in the figure 4. For 8 of 4 categories × 8 size GT map, 4 times down-sampling, output as 4 channel probability map, each probability map contains the original 4 × 4. Regional probability information. The output calculation formula is as follows:

$$Y'_{k(l,m)} = \frac{1}{s^2}\sum_{(i,j)} u_{i,j}, with\ u_{i,j} = \begin{cases} 1 & \text{if } Y_{(i,j)}=k \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

In the formula, l, m is the size of the image after down-sampling, i, j is the pixel coordinates before down-sampling, k is the category corresponding to the classification task, and s is the down-sampling magnification. $u_{i,j}$ is the pixel value corresponding to the output image , given a down-sampling ratio s and the size of the down-sampling characteristic image (width l, height m), the CPP pooling calculation range of the input image $Y_{(i,j)}$ is $s \times l \le i < s \times (l+1), s \times m \le j < (m+1)$ .CPP pooling will calculate the probability of each class in the corresponding calculation area, and then normalize the output graph.

In this way, after each down-sampling of the network encoder segment, the auxiliary classifier will obtain a rough semantic segmentation probability map and a probability map of the GT tag obtained after using CPP pooling. At this time, the auxiliary loss function L_{cpp} can be designed to add to the main segmentation network loss function to optimize the network down-sampling results. Because the U-Net structure encoder segment has four down-sampling, there are four auxiliary classifiers in total, and the size of the network GT tag graph is 1/2, 1/4, 1/8, 16/1 respectively, so the corresponding down-sampling ratio of the Cpp pooling layer is 2, 4, 8, 16.

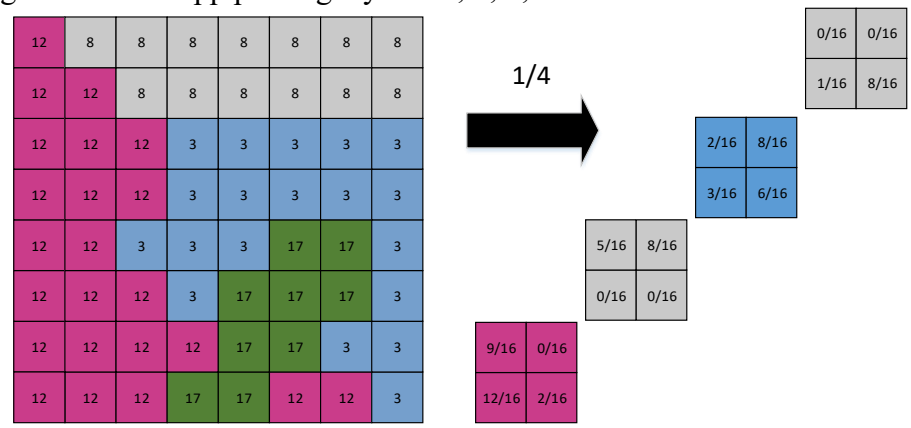

Figure 4: CPP pooling example, 8 × 8 Label graph for 4 times lower sampling.

3. Experiment and Result Analysis

3.1. Experimental Data Set

Data set uses two data sets, data set 1[13] has 20 COVID-19 patients' CT data, and data set 2[14] has 50 COVID-19 patients' data. The file is in NIFTI format, and its data format is converted to 2D CT image, and the size is cut to 256 × 256, and the non-infected images were removed to obtain 2629 CT images. In the CT label image, the gray value of the focus area of COVID-19 is 1, the gray value of the background area is 0, and the training image is a gray image with a gray value of 0~255. The data set is divided into training set, verification set and test set according to the ratio of 8:1:1. 2104 pictures of training set, 262 pictures of verification set and 263 pictures of test set were obtained， and explore a better context information modeling mechanism to replace the CBAM module to improve the context information capture capability of the network.

3.2. Experimental Environment and Parameter Configuration

The experimental environment uses the CUDA 10.0 platform to speed up the training. The graphics card model is TITAN Xp, and the size of 12G video memory. The model is built using torch1.3.1, the learning rate parameter size is 0.001, the weight attenuation parameter size is 0.0001, and the optimization function is the RMSProp algorithm. The loss function of the network uses the sum of the two-class cross-entropy loss and the loss function of the auxiliary classifier, and the number of training epochs is 70. The figure 5 shows the loss function curve of this experiment. The model training loss and the verification loss tend to be consistent and will not decline.

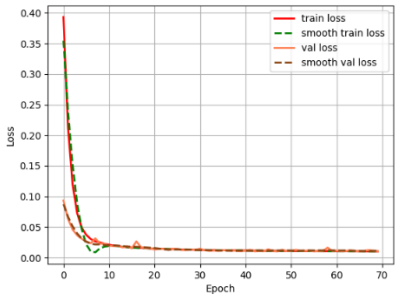

Figure 5: Loss function curve.

3.3. Loss Function

The loss function used in the experiment is the loss of binary cross entropy of the backbone segmentation network and the L_{cpp} loss of auxiliary classifier. The calculation formula of the L_{cpp} loss of the GT label graph probability and the rough segmentation probability graph for the down-sampling ratio s is as follows:

$$D_{KL}\left(p(x)\|q(x)\right)=-\sum_{x}p(x)\log\frac{q(x)}{p(x)}$$
$$L_{cpp}=\frac{1}{H\times W}\sum_{h,w\in H,W}D_{KL}\left(p(x)\|q(x)\right) \tag{3}$$

Where, $p(x)$ is the probability distribution of the pixel position corresponding to the rough

segmentation semantic probability map, and $q(x)$ is the probability distribution of the pixel position corresponding to the GT tag map using Cpp pooling. The total network loss function is:

$$L = \sum_{s \in S} L_{cpp} + BCELoss \tag{4}$$

3.4. Ablation Experiment

We conducted the ablation experiment with the original U-Net network as the benchmark to verify the segmentation performance of the multi-scale attention module added to the U-Net, as well as the segmentation performance of the multi-scale attention module and auxiliary classifier added to the U-Net. The results of the ablation experiment are shown in Table 1. From the experimental data, there is a big gap between the basic U-Net and our proposed model. Especially in the segmentation accuracy and recall rate of the focus area, our model has improved greatly, and other indicators have improved compared with U-Net. The experiment proves that the module used in our model is useful.

Table 1: Results of ablation experiment.

Baseline	Multi-scale attention module	Auxiliary classifier	Precision	Recall	IoU	MIoU
√			83.21	76.52	71.01	82.07
√	√		85.62	80.33	74.87	85.23
√	√	√	**87.22**	**84.55**	**75.98**	**86.82**

3.5. Network Comparative Analysis

In order to test the difference in performance between this network model and other medical segmentation models, U-Net, FCN[15], Mask-RCNN[16] and this network model were used for comparative experiments in the same experimental data set. The accuracy of the model is 4.01%, 6.43% and 3.79% higher than that of U-Net, FCN and Mask-RCNN, and the recall rate is 8.33%, 10.93% and 3.09% higher respectively. It is proved that the network proposed in this paper is superior to the traditional segmentation network in terms of various segmentation indicators. The detailed results of the comparative experiment are shown in Table 2 below.

Table 2: Comparison of network segmentation results.

Net Model	IoU	Precision	Recall	MIoU
U-Net	71.01	83.21	76.52	82.07
FCN	69.70	80.79	73.62	79.48
Mask-RCNN	72.19	83.23	81.46	84.30
Our method	**75.98**	**87.22**	**84.55**	**86.82**

4. Conclusion

U-Net is a deep learning network commonly used in medicine, but its ability to segment the complex CT image focus area is still limited. The maximum pooling it uses has the disadvantage of information loss and the network itself does not capture multi-scale features, which is extremely important for segmentation. In this paper, a multi-scale attention module based on MSCA structure is used, and CBAM is used to increase the attention of the module in the image space domain, improve the ability of U-Net to model multi-scale information. At the same time, a simple auxiliary classifier is used to obtain the probability distribution map of rough semantic segmentation for the maximum

pooled information loss of U-Net, The probability distribution diagram of GT tag graph is obtained by down-sampling the GT tag graph using the class probability reservation pooling. The KL divergence of the two is calculated as the auxiliary loss function added to the loss function of the main partition network, so as to minimize the information loss and optimize the classification performance. In the future, auxiliary classifiers can be further designed to improve the classification ability of the network.

References

[1] Shelhamer, Evan et al. "Fully convolutional networks for semantic segmentation." 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014): 3431-3440.

[2] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation//International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2015: 234-241.

[3] Takikawa, Towaki et al. "Gated-SCNN: Gated Shape CNNs for Semantic Segmentation." 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2019): 5228-5237.

[4] Fan D P, Zhou T, Ji G P, et al. Inf-Net: Automatic COVID-19 Lung Infection Segmentation from CT Images. IEEE Transactions on Medical Imaging, 2020, 39(8): 2626 – 2637.

[5] Zhu Lanyun et al. "Learning Statistical Texture for Semantic Segmentation." 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021): 12532-12541.

[6] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.

[7] Szegedy, Christian et al. "Going deeper with convolutions." 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014): 1-9.

[8] Sun Ke et al. "Deep High-Resolution Representation Learning for Human Pose Estimation." 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019): 5686-5696.

[9] Guo M H, Lu C Z, Hou Q, et al. Segnext: Rethinking convolutional attention design for semantic segmentation. arXiv preprint arXiv:2209.08575, 2022..

[10] Peng Chao et al. "Large Kernel Matters-Improve Semantic Segmentation by Global Convolutional Network." 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017): 1743-1751.

[11] Zhao Hengshuang et al. "Pyramid Scene Parsing Network." 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016): 6230-6239.

[12] Han D, Yoo J, Oh D. See Through Net: Resurrection of Auxiliary Loss by Preserving Class Probability Information//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 4463-4472.

[13] Ma J, Wang Y, An X, et al. Towards Efficient COVID-19 CT Annotation: A Benchmark for Lung and Infection Segmentation. arXiv, 2020.

[14] Morozov S P, Andreychenko A E, Pavlov N A, et al. MosMedData: Chest CT Scans with COVID-19 Related Findings Dataset. 2020.

[15] Shelhamer Evan et al. "Fully convolutional networks for semantic segmentation." 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014): 3431-3440.

[16] He Kaiming et al. "Mask R-CNN." 2017 IEEE International Conference on Computer Vision (ICCV) (2017): 2980-2988.