Medical Image Segmentation of COVID-19 Based on Deep Learning

Junbo Zeng 36920221153144[1], Yuchen Zhang 36920221153151[1], Haipeng Xu 36920221153132[1], Zhengyi Zhang 36920221153152[1] and Xuyang Yu 36920221153143[1]

[1] *Institute of Artificial Intelligence, Xiamen University*

Abstract

COVID-19 is a highly contagious respiratory disease, the cumulative number of infections worldwide exceeds 158 million and is still growing rapidly. CT is one of the important methods for diagnosing COVID-19.The clinical manifestation of CT is an important basis for judging the progress of the disease. With the rapid increase of infected patients, CT data has increased exponentially, which brings a huge burden to doctors' diagnosis work. Therefore, using computer-aided diagnosis technology for COVID-19 segmentation can greatly improve the efficiency of doctors' diagnosis. Since the shape and size variability of the infected region, the deep learning method is more generic than the traditional image segmentation algorithm. The segmentation method of the COVID-19 infection based on deep learning is as follows: First, preprocess the original image to reduce the interference of useless information;Then, the 3D U-Net model is used to process the data and perform three-dimensional modeling of the segmentation results.

Keywords

LaTeX class, paper template, paper formatting, CEUR-WS

1. Introduction

COVID-19 is a highly contagious respiratory disease, the cumulative number of infections worldwide exceeds 158 million and is still growing rapidly. CT is one of the important methods for diagnosing COVID-19. The clinical manifestation of CT is an important basis for judging the progress of the disease. With the rapid increase of infected patients, CT data has increased exponentially, which brings a huge burden to doctors' diagnosis work. Therefore, using computer-aided diagnosis technology for COVID-19 segmentation can greatly improve the efficiency of doctors' diagnosis. Since the shape and size variability of the infected region, the deep learning method is more generic than the traditional image segmentation algorithm.

Medical image segmentation plays an indispensable role in the diagnosis and analysis of the disease, and the segmentation results will affect the subsequent diagnosis of the disease. Early methods of lung CT image segmentation mainly used basic image processing methods, including image operation and morphological processing. With the development of neural network technology, researchers have applied deep learning to medical image segmentation.

At present, medical image segmentation algorithms based on deep learning include full convolutional neural network FCN, U-Net based on encoder and decoder, DeepLab based on hollow convolution, Mask RCNN based on object detection, RefineNet based on feature fusion, adversarial network GAN, etc.

It can be seen from the above deep learning-based method that the segmentation method is to obtain the deep information of the image through convolution and pooling operations, and then learn through the back propagation training. The deep learning method is more like a black box model. The internal parameters are constantly updated through the self-training of the model, the predicted results of the model are compared with the labeled gold standard results, and the updating direction of the internal parameters is determined by the loss function until the value of the loss function no longer decreases, and the training of the model is completed.

2. Related work

Over the past decade, deep convolutional neural networks have been widely used for medical image segmentation and have shown adequate performance. Transformer architectures have been successful in many natural language processing tasks. However, its application in medical vision remains largely unexplored.

Jeya et al. [10]proposed the Gated Axial-Attention model to extend existing architectures by introducing an additional control mechanism in the self-attention module. It is solved that the number of medical imaging data samples is relatively small, making it difficult to efficiently train Transformers for medical applications. Furthermore, a local-global training strategy (LoGo) is proposed, operating on the whole image and patch to learn global and local features, respectively.

Yuanfeng Ji et al. [5] proposed a multi-composite Transformer (MCTrans) model that can be easily plugged into UNet-like networks. MCTrans embeds multi-scale convolutional features as a sequence of tokens and performs intra- and inter-scale self-attention instead of single-scale attention in previous work. Furthermore, a learnable proxy embedding is introduced to model semantic relations and feature augmentation by using self-attention and cross-attention, respectively.

Yunhe Gao et al. [1] proposed UTNet to integrate self-attention into convolutional neural networks to enhance medical image segmentation. UTNet applies a self-attention module in the encoder and decoder to capture long-range dependencies of different scales with minimal overhead. UTNet adopts an efficient self-attention mechanism along with relative position encoding, which significantly reduces the complexity of self-attention operations. Furthermore, UTNet adopts a novel self-attention decoder to recover fine-grained details from connections skipped in the encoder.

Ge-Peng JI et al. [4] proposed PNS-Net (Progressive Normalization Self-Attention Network), Existing video polyp segmentation (VPS) models usually employ convolutional neural network (CNN) to extract features. However, due to their limited receptive fields, CNNs cannot fully exploit the global temporal and spatial information in consecutive video frames, leading to false positive segmentation results. PNS-Net is completely based on the basic normalized self-attention block, fully equipped with recursion and CNN.

Yinglin Zhang et al. [11] proposed a Multi-Branch hybrid Transformer network (MBT-Net) based on Transformer and body-edgebranch, using convolutional blocks to focus on local texture feature extraction, and establishing pairs of spatial, channel and remote dependencies of layers. Use body-edge branches to promote local consistency and provide edge location information.

3. Proposed Solution

3.1. RA-Unet model

In this paper, U-Net is used as the basic backbone network, the encoder extracts the data features, and the decoder recovers the extracted features. The encoder and the decoder use the hopping connection for feature complementation. Combining with the residual block to extract deep semantic information, the penultimate layer of RA embedded decoder is constructed as RA-Unet (Residual attention U-net, RA-Unet) for 3D images.

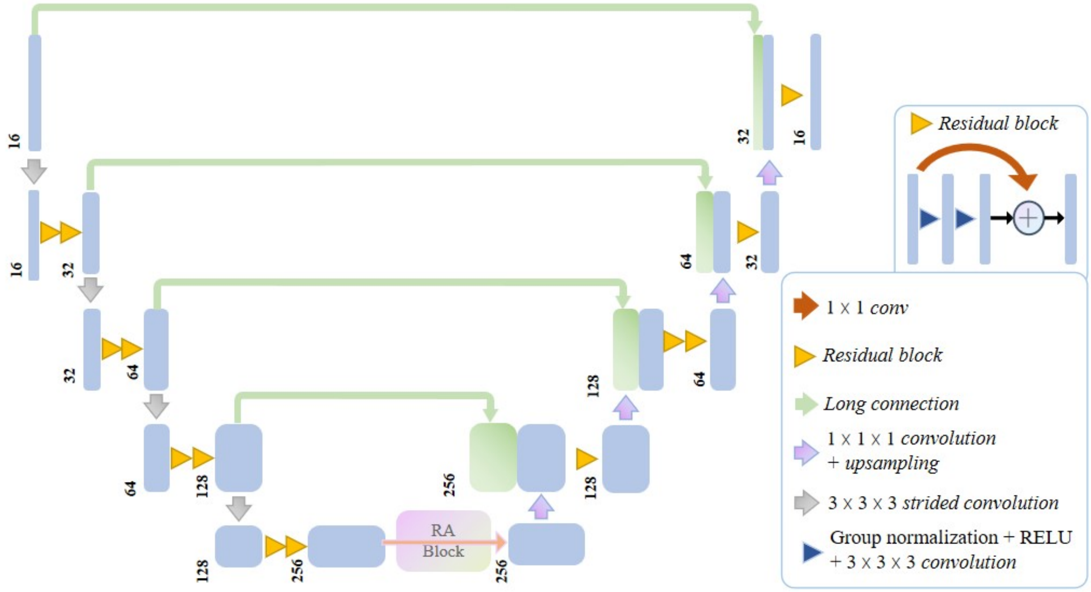


Figure 1: The proposed RA-Unet architecture

As shown in FIG. 1, RA-Unet uses the U-Net model as the basic backbone network, and constructs a 5-layer encoder and decoder respectively. the encoder embeds two residual blocks per layer and the decoder embeds one residual block per layer. The residual block activation function uses a Rectified Linear Unit (ReLU) function. RA is located at the penultimate layer of the decoder path for feature extraction. each decoding layer combines as input the output of the previous decoding layer with the jump-connected output of the encoding layer. The number of input channels for the first layer is 16, and the number of channels is doubled for each layer, in order: 3264128256. The 3×3×3 convolution kernel with step size 2 is used to extract the features between each layer, and the bilinear interpolation is used to restore the features. A 1×1×1 convolution and sigmoid function activation were performed on the output portion of the model . RA is the combination of attention and residual error, which makes the model can select and synthesize in global receptive field and local information according to requirements, so as to improve the accuracy of the model and accelerate the convergence speed of the model.

3.2. Combination of attention and residual block

As the depth of the model increases, more features can be extracted from the model. But at the same time, the model will be difficult to train because of the deepening of the depth, which leads to the degradation of the model. He et al. [3] proposed a residual structure that allows the model to be comprehensively selected between identity mapping and convolution operation, which solves the problem of degradation caused by deepening of the model and improves the performance. In this paper, RA is constructed by combining attention and residual structure, which makes the model perform convolution and attention selection automatically during training. This speeds up model convergence and allows the model to focus on areas of interest. As shown in FIG. 2, RA is obtained by replacing the second layer of the three-layer residual block of the Residual Network 50 (ResNet50) with an attention module [9], and making the output dimension of the last layer the same as the input dimension. The mathematical expression of RA is shown in formula (1):

$$H\left(x\right) = F\left(x\right) + x \tag{1}$$

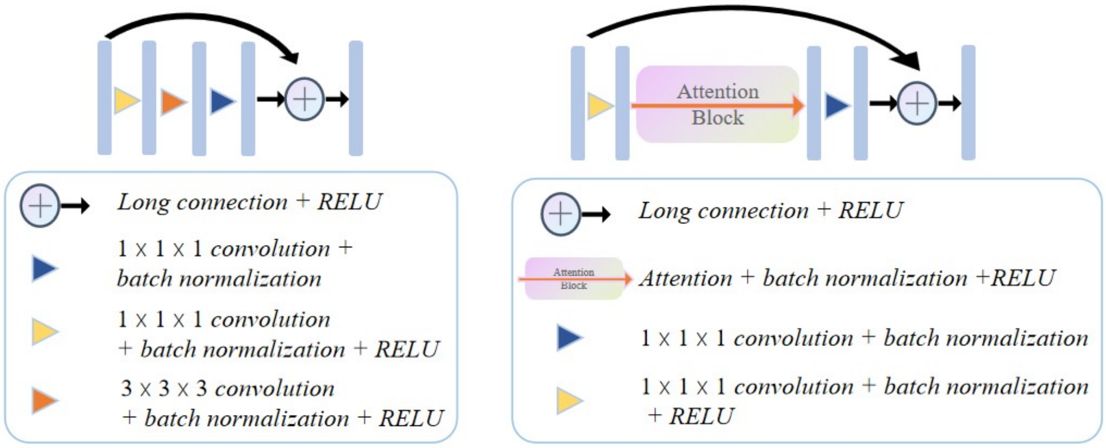
where x is the input, H(.) is the output, and F(.) is the convolution, attention, and nonlinear transformation operations. By incorporating attention into the residual block, the model can have a more flexible structure. During the training process, the model can be convoluted, selected and synthesized with attention. For example, when F(x) approaches 0, model selection inhibits attentional operation; On the contrary, model selection combines attention with convolution operation to synthesize global information and local information. The combination of attention and residual block is beneficial to improve the accuracy of the model and accelerate the convergence.

3.3. attention

The attention mechanism only focuses on one input or output feature for spatial learning, which can effectively capture context information and solve the dependence relationship between learning long-distance features. Tsai et al. [2] interpreted the attention formula as shown in equation (2):

$$A\left(Q,K,V\right) = \sum\nolimits_{i=1}^{L} \frac{k\left(q_i * ki^T\right)}{\sum\nolimits_{j=1}^{L} k\left(q_j,k_j\right)} * V_i = Ep\left(k_i|q_i\right) * \left[V_i\right] \tag{2}$$

where A (·) is the attention formula; The input features are linearly transformed to obtain Q, K and V as the input of attention. q_i and q_j are elements of Q; k_i and k_j are elements of K; $k\left(q_i,k_i\right)$ is $\exp\left(\frac{q_i * k_i^T}{\sqrt{d}}\right)$, where d is the scaling factor; L is the length of attention input, $p\left(q_i|k_i\right) = \frac{k\left(q_i,k_i\right)}{\sum\nolimits_{i=1}^{L} k\left(q_j,k_j\right)}$ Q and K are subjected to similarity calculation to obtain a weight value, and the weight value is normalized and subjected to weighted summation with V to obtain


Long connection + RELU
1 x 1 x 1 convolution + batch normalization
1 x 1 x 1 convolution + batch normalization + RELU
3 x 3 x 3 convolution + batch normalization + RELU
Attention Block
Long connection + RELU
Attention Block
Attention + batch normalization +RELU
1 x 1 x 1 convolution + batch normalization
1 x 1 x 1 convolution + batch normalization + RELU

(a) Residual block

(b) Residual attentional block

Figure 2: Illustration of residual block and RA block

attention. Inspired by Srinivas et al. [8], this paper proposes an attention module, which uses three-dimensional position coding information to enable the model to obtain higher-dimensional context-related information. The proposed attention module is shown in Figure 3. $\oplus$ denotes addition element by element, $\otimes$ represents an input three-dimensional characteristic graph; C, D, H and W respectively represent the channel number, length, height and width of the input X; w_d, w_h and w_w are learnable parameters; P is obtained by adding w_d, w_h and w_w element by element, and is the relative position code; the linear transformation is a convolution operation with a convolution kernel size of 1 * 1 * 1, and X is respectively subjected to three convolution operations to obtain Q, K and V; The output is Z.

4. emperiments

4.1. experimental data

- Paiva, O., 2020. CORONACASES.ORG - Helping Radiologists To Help People In More Than 100 Countries! | Coronavirus Cases - . [online] Coronacases.org. Available at: <link> [Accessed 20 March 2020].
- Glick, Y., 2020. Viewing Playlist: COVID-19 Pneumonia | Radiopaedia.Org. [online] Radiopaedia.org. Available at: <link> [Accessed 20 April 2020].

4.2. data partitioning

- 16 data for training
- 3 data for test

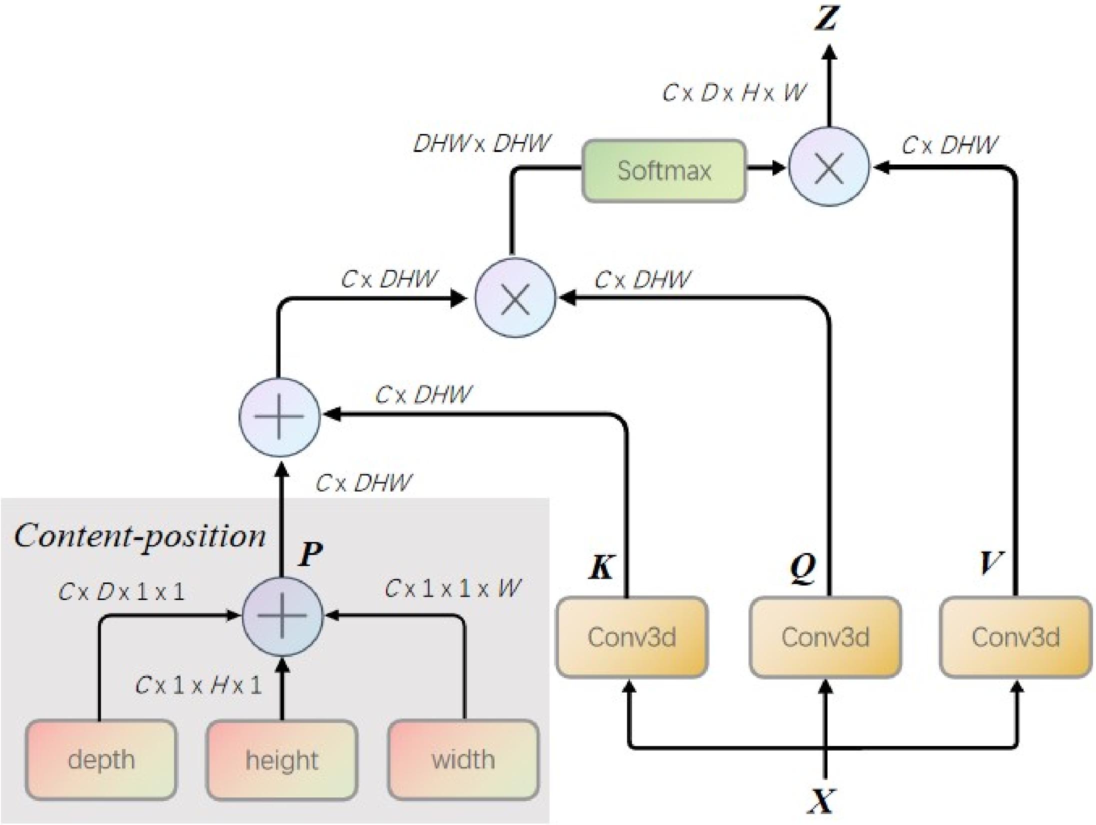


Figure 3: Illustration of attention

4.3. data preprocessing

- numerical truncation :clipped on the min (-1000) and max (+1000) interesting Hounsfield Unit range
- normalization
- Cut: 512*512*301 cut to 192*192*32

4.4. Experimental environment and parameter setting

in that lab in this article, use a configuration that: Community Enterprise Operating System Ubuntu (16.04.5 LTS, Red Hat, United States); Tesla (V100, Nvidia, United States); Python (3.8, Python Software Foundation, United States); pytorch 1.8.1; Gradient descent was performed using an adaptive moment estimation (Adam) optimizer [6] with Dice loss as the loss function [7]. In terms of parameter setting, the Batch size is set to 4, the learning rate is set to 0.0001, and the epoch is set to 70.

4.5. Evaluation indicators

The evaluation index uses Dice similarity coefficient (DSC)(expressed by symbol DSC)]. DSC is a measure of the similarity between two samples and tends to compare the similarity of the padding within two samples. The DSC formula is shown as:

$$Dice\left(P,T\right)=\frac{2\left|P\cap T\right|}{\left|P\right|+\left|T\right|} \tag{3}$$

4.6. Experimental results

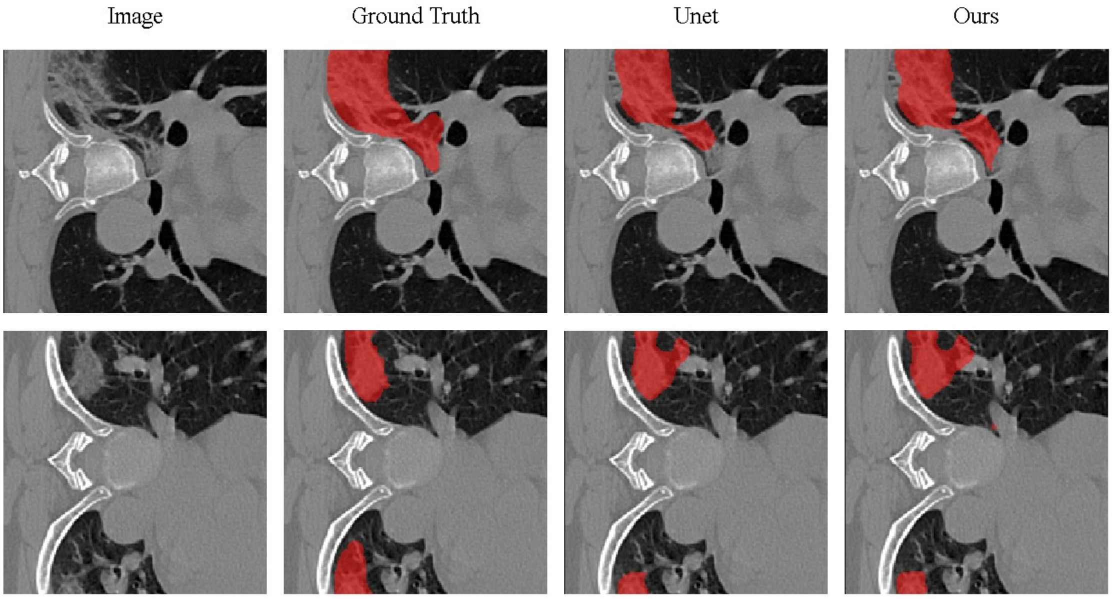


Figure 4: Experimental results

Table 1
Compare of differentmodel prediction results

method	U-net	ours
DSC	0.6641	0.6813

5. Conclusion

CT images of the lungs of patients with neoconiosis can be used as an effective basis for early detection and diagnosis of neoconiosis, and have important analytical research value for the segmentation of the infected region in the images. At present, there are still some difficulties in the segmentation of the infected region of neointimal due to the specificity of the lesion, such as the contrast between the infected region and the lung tissue is not obvious, its grayscale value is not uniform and the boundary is not continuous. In this paper, we use U-Net as the basic backbone network for the segmentation of infected regions of lung CT images of patients with neocrown pneumonia, and use jump connections between encoder and decoder for feature complementation. Combining attention with residual structure to build RA makes the model perform convolution and attention ground selection autonomously during the training process. By incorporating attention into the residual block, it allows the model to have a more flexible structure, and the model can perform convolution, attention ground selection and synthesis during the training process, which in turn improves the model accuracy and accelerates the model convergence speed. In addition, this paper also proposes an attentional module that uses three-dimensional location encoding information so that the model can obtain higher dimensional contextual association information.

References

[1] Yunhe Gao, Mu Zhou, and Dimitris N. Metaxas. “UTNet: A Hybrid Transformer Architecture for Medical Image Segmentation”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021.* Ed. by Marleen de Bruijne et al. Cham: Springer International Publishing, 2021, pp. 61–71.

[2] Jun Han and Claudio Moraga. “The influence of the sigmoid function parameters on the speed of backpropagation learning”. In: *International workshop on artificial neural networks.* Springer. 1995, pp. 195–201.

[3] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2016, pp. 770–778.

[4] Ge-Peng Ji et al. “Progressively Normalized Self-Attention Network for Video Polyp Segmentation”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021.* Ed. by Marleen de Bruijne et al. Cham: Springer International Publishing, 2021, pp. 142–152.

[5] Yuanfeng Ji et al. “Multi-compound Transformer for Accurate Biomedical Image Segmentation”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021.* Ed. by Marleen de Bruijne et al. Cham: Springer International Publishing, 2021, pp. 326–336.

[6] A KingaD. “A methodforstochasticoptimization”. In: *Anon. InternationalConferenceon Learning Representations. SanDego: ICLR* (2015).

[7] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. “V-net: Fully convolutional neural networks for volumetric medical image segmentation”. In: *2016 fourth international conference on 3D vision (3DV).* IEEE. 2016, pp. 565–571.

[8] Aravind Srinivas et al. “Bottleneck transformers for visual recognition”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 2021, pp. 16519–16529.

[9] Yao-Hung Hubert Tsai et al. “Transformer Dissection: A Unified Understanding of Transformer’s Attention via the Lens of Kernel”. In: *arXiv preprint arXiv:1908.11775* (2019).

[10] Jeya Maria Jose Valanarasu et al. “Medical Transformer: Gated Axial-Attention for Medical Image Segmentation”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021.* Ed. by Marleen de Bruijne et al. Cham: Springer International Publishing, 2021, pp. 36–46.

[11] Yinglin Zhang et al. “A Multi-branch Hybrid Transformer Network for Corneal Endothelial Cell Segmentation”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021.* Ed. by Marleen de Bruijne et al. Cham: Springer International Publishing, 2021, pp. 99–108.