ISSN 2616-5775 Vol. 6, Issue 8: 1-7, DOI: 10.25236/AJCIS.2023.060801

Method of rectal tumor segmentation based on ResUnet++

Mingao Liu*

School of Computer and Software Engineering, University of Science and Technology Liaoning, Anshan, 114000, China
**Corresponding author: 1249277436@qq.com*

Abstract: *Rectal cancer is one of the most common malignant tumors. Electronic cross section examination (CT) is used as a screening tool in the diagnosis of rectal cancer. The application of computer aided diagnosis technology to help doctors distinguish between benign and malignant tumors in rectal CT images is of great significance to guide further clinical treatment. In this paper, we analyze the performance of the current mainstream neural network models using the rectal tumor data set from the 7th Teddy Cup Data Mining Challenge B. Among them, ResUnet ++ achieves Dice value of 83.32% and IoU value of 70.06%, which is the best performance among mainstream models.*

Keywords: *rectal cancer, CT images, neural network models*

1. Introduction

With the improvement of China's economy, people's diet has changed significantly. In China, colorectal cancer has attracted much attention as one of the most common malignant tumors. Rectal cancer refers to the malignant tumor within 15 cm from the anal margin [1]. The mass in the rectal area and the swelling of the surrounding lymph nodes can be observed in CT images, which may block the intestine. Even if the medical professionals have rich clinical experience, long working hours lead to excessive fatigue, which inevitably leads to a certain degree of misdiagnosis. Therefore, accurate segmentation of rectal tumors is crucial in the subsequent diagnosis and treatment of colorectal cancer. At present, computer technology has been widely concerned in assisting CT image segmentation. Using modern information technology, we built a deep learning-based tumor recognition and segmentation model for rectal cancer images, which greatly reduced the work burden of medical experts and saved a lot of time for subsequent diagnosis and treatment, and has extremely important practical significance.

2. Correlation model

Traditional convolutional neural networks extract features through the convolutional layer and output results through the fully connected layer, but the results are all a class probability value of the input image. This network structure is very suitable for image classification tasks. However, the task of image segmentation is different from the task of classification, the key point of the task is how to accurately classify each pixel in the image to achieve the purpose of segmentation. In order to realize image segmentation technology, Long[2] et al proposed full convolutional neural network (FCN). For medical images, the FCN network does not take into account the global context information, resulting in unsatisfactory segmentation details. In order to solve the above problems, Ronneberger[3] et al. proposed the Unet network model. Although the Unet network model has achieved good results in medical image segmentation, due to its internal use of cascade structure, similar low-level features will be repeatedly extracted, and the model parameters and computing resources will be overused, resulting in unsatisfactory segmentation results. In order to solve the above problems, Oktay[4] et al. proposed the Attention Unet network model, which added attention blocks on the basis of Unet network model. This attention block can inhibit the feature activation of irrelevant regions in the model through the internal attention mechanism, which improves the segmentation ability of the model. In 2018, Zhou[5] et al. proposed the Unet++ network model, which was improved on the original Unet[3] network model. First, the Unet++ network performs an upsample after each downsample of the Unet network. The Unet++ network model is based on the encoder-decoder structure, and since the Unet++ network has redesigned the skip path connection, its encoder and decoder connection has also changed. When the feature

ISSN 2616-5775 Vol. 6, Issue 8: 1-7, DOI: 10.25236/AJCIS.2023.060801

mapping between the encoder and the corresponding decoder is semantically similar, the performance of the model will be improved. ResUnet, proposed by Zhang[6] et al., is a very successful deep convolutional neural network structure with strong feature expression ability and shallow network depth, which enables it to show excellent performance in tasks such as image classification. The ResUnet ++ model proposed by Debesh et al. [7] utilizes residual block, SE module, ASPP module and attention block. The residual blocks propagate information across the layers, allowing a deeper neural network to be built that can solve the degradation problem (the problem of disappearing gradients) in each encoder. This improves channel interdependence while reducing computational costs.

3. ResUnet ++ model

The ResUnet ++ architecture is a semantically segmented neural network using residual blocks, SE modules, ASPP modules and attention blocks. The residual blocks propagate information across the layers, allowing a deeper neural network to be built that can solve the degradation problem (the problem of disappearing gradients) in each encoder. This improves channel interdependence while reducing computational costs. The ResUnet ++ architecture consists of a stem block followed by three encoder blocks, ASPP, and three decoder blocks. The architecture framework of ResUnet ++ is shown in Figure 1.

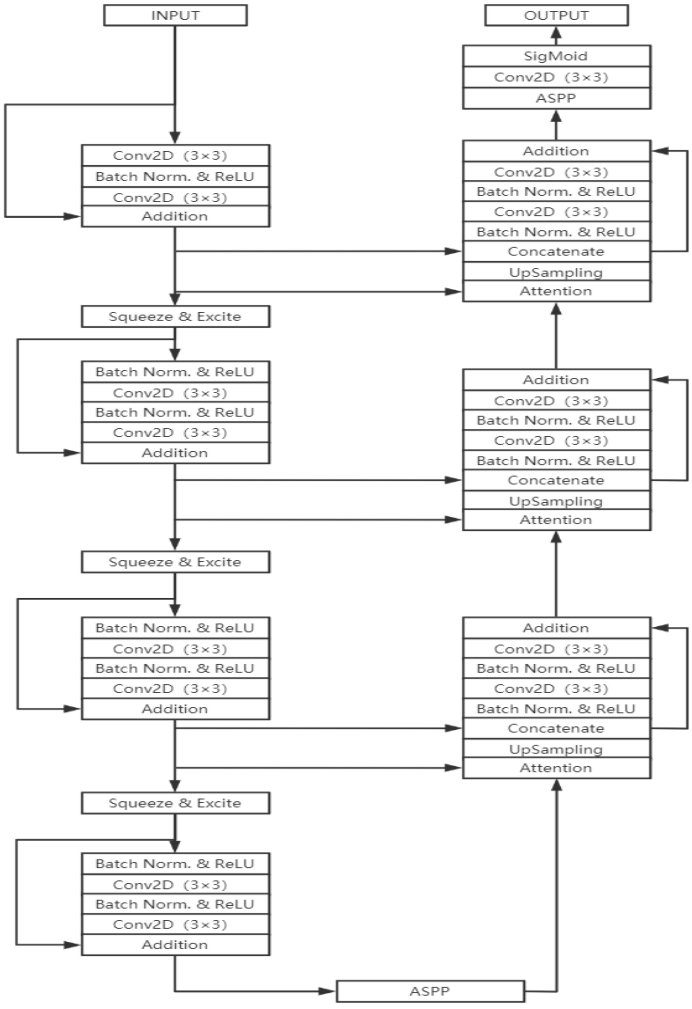


Figure 1: ResUnet ++ architecture

Each encoder block consists of two consecutive 3x3 convolution blocks and a unit map, each convolution block consisting of a BN layer, a Relu layer and a convolution layer. In the first convolution layer of encoder block, the step convolution layer is applied to reduce the spatial dimension of feature mapping by half. The output of the encoder block is transmitted through the SE module [8]. ASPP[9] acts as a bridge, broadening the view of the convolution kernel to include a broader context. The decoder counterpart is also composed of residual units, and before each decoder unit, the attention module

ISSN 2616-5775 Vol. 6, Issue 8: 1-7, DOI: 10.25236/AJCIS.2023.060801

increases the effectiveness of the feature mapping. The next step is to up-sample the lower-level feature map and connect it with the feature map of the corresponding encoding path. The output of the decoder is connected through the ASPP module, and finally the convolution is activated by 1x1 Sigmod to obtain the segmentation graph.

3.1. Unet++ network model

ResUnet++ is modified on the basis of the Unet ++ model. The Unet++ network model is also based on the encoder-decoder structure, and since the Unet++ network has redesigned skip path connections, its encoder and decoder connections have also changed. For Unet networks, the encoder part passes the feature information to the decoder part by skipping the path connection. However, in Unet++, the feature information extracted by the encoder is first transmitted to the dense convolution block in the middle for feature extraction, and finally to the decoder part. At the front of each convolution layer is the connection layer, the main function of the connection layer is to fuse the output of the previous convolution of the current dense convolution block with the output of the upper sample of the lower dense convolution block. In essence, the main function of dense convolution blocks is to reduce the semantic gap between encoder and decoder feature mappings. When the feature mapping between the encoder and the corresponding decoder is semantically similar, the performance of the model will be improved. The network model of Unet++ is shown in Figure 2.

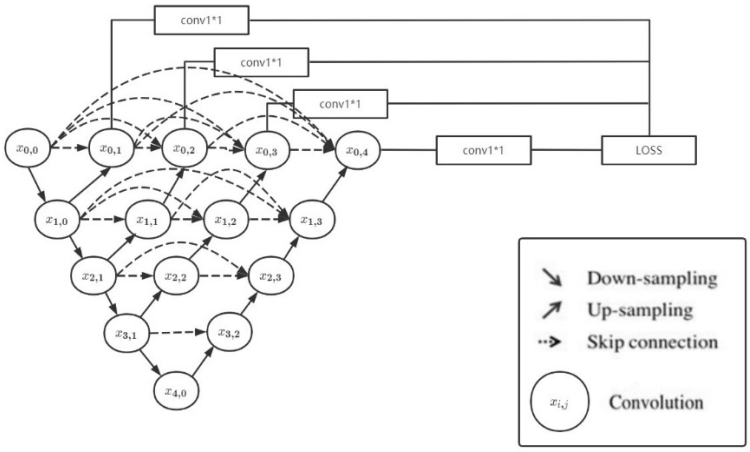


Figure 2: Unet++ network structure

Taking the first layer of the network as an example, Figure 3 is the schematic diagram of the input and output formula of the first layer.

$$x_{0,1} = H[x_{0,0}, U(x_{1,0})]$$

$$x_{0,2} = H[x_{0,0}, x_{0,1}, U(x_{1,1})]$$

$$x_{0,3} = H[x_{0,0}, x\mathbf{0}, \mathbf{1}, U(x_{1,2})]$$

$$x_{0,4} = H[x_{0,0}, x_{0,1}, x_{0,2}, x_{0,3}, U(x_{1,3})]$$

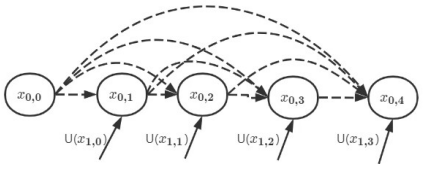


Figure 3: Input and output diagram of the first layer

Let $X_{i,j}$ be the output of nodes $x_{i,j}$ where i represents sequential search for the down-sampling layer in the encoder direction and j represents sequential search for the convolution layer of the dense block in the skip path direction.The output $X_{i,j}$ is calculated as follows:

$$X_{i,j} = \begin{cases} H(x_{i-1,j}, j), j = 0 \\ H([[x_{j-1,0}], U(x_{j+1,j-1})]), j > 0 \end{cases} \quad (1)$$

The function H(•) is a convolution operation on the input image, and there is an activation function

ISSN 2616-5775 Vol. 6, Issue 8: 1-7, DOI: 10.25236/AJCIS.2023.060801

after the convolution layer of each layer. The function U(•) represents the up-sampling operation. [] Indicates the connection layer. If the index value of a node j is 0, then the node has only received one input in the first layer before the encoder. If the index value j of a node is 1, then the node receives only two inputs, both of which come from the subnetwork of the encoder. If the index value of a node j is greater than 1, then the node receives j+1 inputs. The j inputs here are the output of the previous node in the same skip path, and the other input is the upsampled output of the skip path below that node. This way, each previous feature map is saved and goes directly to the current node.

The Unet++ network is equivalent to connecting the Unet networks at layers 1 to 4 by means of long and short connections. This connection can extract all the depth features, so that the network can self-learn the important features in the image.

3.2. Residual unit

Deep neural networks are relatively challenging to train. As the depth of the network increases, training deep neural networks can improve accuracy. But at the same time too deep network will cause the problem of gradient disappearance. Residual units [10] make the network easy to train, and residual connections within the cells help propagate information without degradation. The residual network is proposed to solve the degradation problem of deep neural network (DNN) when there are too many hidden layers. Degradation refers to the problem that when there are more hidden layers of the network, the accuracy of the network reaches saturation and then deteriorates sharply, and this degradation is not caused by overfitting. This intriguing hypothesis inspired Dr. He, who proposed residual learning to solve the problem of degradation. For a stack structure (made up of several layers), the features $H(x)$ it learns when the input is x, now we want it to learn residuals $F(x) = H(x) - x$, so that the original learning features are $F(x) + x$. The reason for this is that learning residuals is easier than learning raw features directly. When the residual is 0, then the stack layer only does the identity mapping, at least the network performance will not deteriorate, in fact, the residual will not be 0, which will also make the stack layer learn new features based on the input features, so as to have better performance. This is somewhat similar to a "short circuit" in an electrical circuit, so it is a short-circuit connection. The residual learning unit is shown in Figure 4.

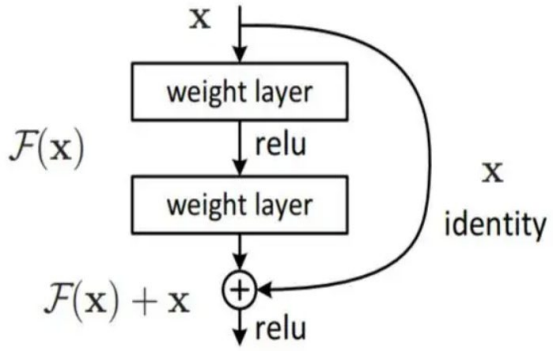


Figure 4: Residual learning unit

Why residuals learning is relatively easier? Intuitively, there is less to learn in residuals, because residuals are generally small and learning is less difficult. However, we can analyze this problem from a mathematical point of view. First, the residual unit can be expressed as:

$$y_l = h(x_l) + F(x_{l,}\ W_l) \tag{2}$$

$$x_{l+1} = f(y_l) \tag{3}$$

Where and represent the input and output of the fourth residual unit, respectively, noting that each residual unit generally contains a multi-layer structure. Is the residual function, which represents the learned residual, and represents the identity mapping, which is the ReLU activation function. Based on the above formula, we obtain the learning characteristics from shallow layer to deep layer as follows:

$$x_L = x_l + \sum_{i=l}^{L-1} F(x_i, W_i) \tag{4}$$

Using the chain rule, the gradient of the reverse process can be found, and the formula can be found to show that the gradient does not disappear. So residual learning will be easie

ISSN 2616-5775 Vol. 6, Issue 8: 1-7, DOI: 10.25236/AJCIS.2023.060801

4. Experimental results and analysis

All the data set used in the experiment is a generic data set, namely, arterial phase and venous phase B. The data set is divided into two sets of CT images, arterial phase and venous phase. We selected the CT image of arterial phase as the sample data of our experiment. A total of 3029 images and corresponding image masks are included, including 2169 positive samples and 860 negative samples. The data size of the image sample is 512*512 gray scale, and its data format is dicom format.

Since deep learning requires the support of large data sets, the data sets need to be extended. Now the original data set is flipped horizontally, rotated 45°, 90°, 180° and other methods to expand the data set to 5 times the original, a total of 15145 frames. The sample data set is divided into training set, verification set and test set according to the ratio of 6:2:2. Among them, the training set samples are 9087, the verification set samples are 3029, and the test set samples are 3029.

4.1. Experimental setup

(1) Experimental hyperparameter setting

In this experiment, the number of cycles is set to 80 times, the batch size is set to 6, the optimizer is Adam, the learning rate is 0.0002, and the loss function adopts the mean square error function. We can adjust the learning rate through the ReduceLROnPlateau callback function provided by the system. If the loss value of the verification set is not reduced for 20 consecutive cycles, the learning rate will become 1/10 of the original.

(2) Data preprocessing

Preprocessing is also one of the more common methods in the field of medical image processing. Preprocessing can reduce the impact of image brightness, contrast and other attributes on the image, and it can also increase the proportion of important information in the overall content of the image, and can simplify the image data, thus saving the time used in the later processing work. The pre-processed image can get the cutting result close to the actual, so the image pre-processing operation must be done before the formal image segmentation.

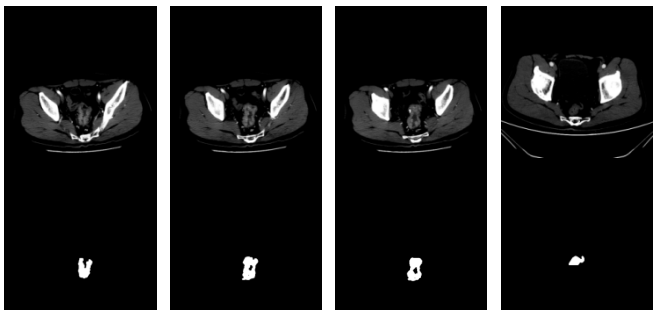


Figure 5: Sample images of four groups of rectal tumors

As shown in Figure 5, the first act is the original image, and the second act is the real mask of the original image. The sample data size of the CT image is 512*512 gray scale. In order to reduce the requirements for computer video memory, the original image size is first adjusted to 256*256 according to our image slicing technology using Python, and then the pixel value of the image is normalized to the range of [0, 255]. Finally, it can be seen from chapter 3 that the image is still processed by window.

4.2. Evaluation index

The performance evaluation of medical image segmentation algorithm is very important in medical image segmentation technology. At present, there are two main evaluation methods: subjective evaluation and objective evaluation. Subjective evaluation refers to the subjective judgment of the accuracy of segmentation results by clinicians with professional knowledge. However, this evaluation method has high requirements for clinicians, has subjective influence, and may change due to environmental differences, so it cannot give an accurate quantitative evaluation. The objective evaluation method compares the similarity between the segmentation result and the real mask image by mathematical algorithm. According to extensive literature research, Dice coefficient and IoU coefficient are commonly used objective evaluation methods. Therefore, these two coefficients were selected as

ISSN 2616-5775 Vol. 6, Issue 8: 1-7, DOI: 10.25236/AJCIS.2023.060801

evaluation indicators in the rectal tumor segmentation experiment in this paper, and their specific definitions and formulas are as follows:

(1)Dice coefficient: Dice coefficient is used to represent the similarity between the image (Y) of the real mask and the predicted output image (Y '), and is usually used to describe the coincidence degree between the segmentation image and the real mask. The calculation formula is as follows:

$$Dice = \frac{2*|Y \cap Y'|}{|Y'|+|Y|} = \frac{2*TP}{FP+FN+2*TP} \tag{5}$$

(2)IoU coefficient: The IoU coefficient represents the proportion of intersection and union between the real mask image and the predicted output image, and is also used to describe the degree of coincidence between the segmentation image and the real mask image. The calculation formula is as follows:

$$IoU = \frac{|Y \cap Y'|}{|Y \cup Y'|} = \frac{TP}{FP+FN+TP} \tag{6}$$

TP represents the number of patients accurately identified as positive. TN refers to true negative, indicating the number of patients accurately identified as negative. FP refers to a false positive and represents the number of patients incorrectly identified as positive. FN refers to a false negative and represents the number of patients who are incorrectly identified as negative.

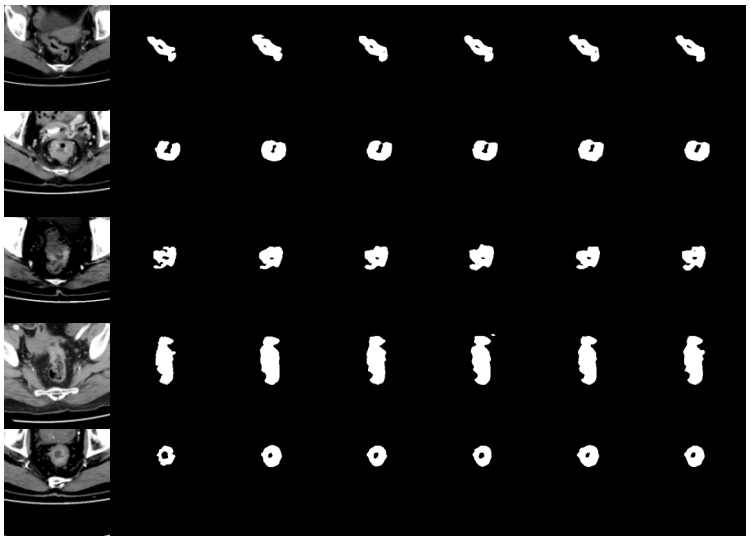
4.3. Analysis of experimental results

We conducted a comprehensive experiment to evaluate the five most advanced deep learning models of Unet[3], Unet++[5], Attention Unet[4], ResUnet[7], Resunet ++[8], respectively, in terms of their segmentation effect and complexity.

Table 1 shows the comparison of different network models on rectal tumor data sets. It can be seen from the data in the table that ResUnet++ has achieved the best performance in terms of Dice value and IoU value.

Table 1: Comparison of different network models on a rectal tumor dataset

Methods	Dice	IoU
Unet[3]	78.19%	61.31%
Unet++[5]	80.56%	66.22%
Attention Unet[4]	81.51%	67.88%
ResUnet[6]	79.94%	65.14%
ResUnet++[7]	82.39%	68.12%



(a) (b) (c) (d) (e) (f) (g)

Figure 6: Qualitative segmentation of sample data by different deep learning network model

In this paper, five groups of sample data are randomly selected. Figure 6 shows the qualitative segmentation results of different deep learning network models for sample data.

Among them, column a represents the original image, column b represents the real mask of the

ISSN 2616-5775 Vol. 6, Issue 8: 1-7, DOI: 10.25236/AJCIS.2023.060801

original image, and column c to g represents Unet[3], Unet++[5], Attention Unet[4], ResUnet[6], ResUnet ++[7] segmentation images respectively. It can be seen from the figure that, compared with other networks, the segmentation result of ResUnet ++ is closer to the real mask image of the original image.

5. Peroration

Aiming at the current mainstream neural network model in rectal cancer segmentation task, this paper analyzes the performance of the current mainstream neural network model using the rectal tumor data set of the 7th "Teddy Cup" data mining Challenge B. Among them, Resun ++ achieves 83.32% Dice value and 70.06% IoU value, which is the best performance among the mainstream neural network models. The effectiveness and performance of Resun ++ model in colorectal cancer segmentation task were proved. In the next step, we will consider improving the Resun ++ model to make it have better performance in medical image segmentation tasks.

Acknowledgement

Liaoning University of science and technology innovation and entrepreneurship training project fund. Project number: S202210146075.

References

[1] Ajit Marecik, Slawomir J, et al. Oncologic and Clinicopathologic Outcomes of Robot-Assisted Total Mesorectal Excision for Rectal Cancer [J]. Diseases of the Colon & Rectum, 2015, 58(7):659-667.
[2] Long J, Shelhamer E, Darrell T. Fully Convolutional Networks for Semantic Segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 39(4):640-651.
[3] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation[C]. 18th International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI 2015, 2015, 9351(1):234-241.
[4] Oktay O, Schlemper J, Folgoc L L, et al. Attention U-Net: Learning Where to Look for the Pancreas [J]. 2018. https://arxiv. org/abs/1804. 03999.
[5] Zhou Z, Siddiquee M, Tajbakhsh N, et al. UNet++: A Nested U-Net Architecture for Medical Image Segmentation [J]. 4th Deep Learning in Medical Image Analysis (DLMIA) Workshop, 2018, 11045:3-11.
[6] Zhang Z, LiuQ, Wang Y. Road Extraction by Deep Residual U-Net [J]. IEEE, 2018(5). DOI: 10. 1109/ LGRS. 2018. 2802944.
[7] Jha D , Smedsrud P H , Riegler M A , et al. ResUNet++: An Advanced Architecture for Medical Image Segmentation[J]. arXiv, 2019. DOI:10. 1109/ISM46123. 2019. 00049.
[8] Hu J, Shen L, Sun G. Squeeze-and-Excitation Networks [C] //2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2018. DOI:10. 1109/CVPR. 2018. 00745.
[9] Wang J, Lv P, Wang H, et al. SAR-U-Net: Squeeze-and-excitation block and atrous spatial pyramid pooling based residual U-Net for automatic liver segmentation in Computed Tomography. 2021[2023-06-28]. CMPB. 2021. 106268.
[10] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition [J]. IEEE, 2016. CVPR. 2016. 90.