Transactions on Materials, Biotechnology and Life Sciences
ISSN: 2960-1932, eISSN: 2960-2319 | Volume 5
MMTCS 2024

Optimisation of U-Net Semantic Segmentation Model Based on Residual Connectivity and Attention Mechanisms

Haosen Jia †, Changrui Zuo †, Zhipeng Zou #, * and Chen Chen #

DUT School of Software Technology & DUT-RU International School of Information Science and Engineering, Dalian University of Technology, Dalian, China

† These authors are co-first authors of the article

# These authors are co-second authors of the article

* Corresponding Author Email: 18563571875@mail.dlut.edu.cn

Abstract. In this paper, an improved U-Net semantic segmentation model based on residual connection and attention mechanism is proposed, aiming to improve the accuracy and stability of image segmentation. Adding residual connections between each convolutional layer of U-Net enables the input to bypass one or more convolutional layers and sum up with the outputs of these layers. This not only improves the training efficiency of the network, but also enhances the robustness and stability of the model. In addition, the model introduces two attention mechanisms based on U-Net: Channel Attention and Spatial Attention. Experimental results on the knee MRI dataset show that the improved model outperforms the traditional U-Net and other comparative methods in evaluation metrics such as loss rate, mean absolute error (MAE), and F1 score, demonstrating its advantages in medical image segmentation tasks. The important information in the image is captured more effectively and the segmentation accuracy is improved. It has important reference value for the medical image processing field.

Keywords: residual connectivity; attention mechanisms; semantic segmentation.

1. Introduction

Semantic segmentation is a technique that is widely used in many fields such as medical imaging, virtual reality, industrial inspection, etc. However, it still has some challenges in handling high resolution images, boundary clarity and multi-scale object recognition. Based on the above-mentioned problems, this paper proposes an improved method based on the traditional U-Net and constructs an improved U-net semantic segmentation model based on residual connection and attention mechanism. By introducing the residual connection, the model effectively solves the gradient vanishing and gradient explosion problems in the deep network and enhances the training efficiency and performance. The combination of the channel attention mechanism and the spatial attention mechanism enables the model to adaptively adjust the weights of the feature maps to capture the key features and details in the image more accurately. The model provides new ideas and methods for the semantic segmentation task.

2. Improved U-Net Semantic Segmentation Model

2.1. U-Net Semantic Segmentation Model

In recent years, with the rise of full convolutional neural networks (FCN) in the field of image segmentation, semantic segmentation technology has made significant progress. However, FCN still faces some challenges in processing high-resolution images, boundary clarity, and multi-scale object recognition.

In this paper, we propose an improved method based on the traditional U-Net and construct an improved U-net semantic segmentation model based on residual connectivity and attention mechanism. The model introduces two attention mechanisms based on U-Net: Channel Attention and

Spatial Attention. By combining these two attention mechanisms, the important information in the image can be captured more effectively and the segmentation accuracy can be improved.

2.2. Overall Network Structure

In this paper, an improved U-Net neural network based on residual connection and attention mechanism is proposed, which combines residual connection and attention mechanism based on traditional U-Net to enhance the training effect and performance of deep neural network. The network structure is specifically divided into an encoder part, a bottleneck layer and a decoder part, and a jump connection is used between the encoder and the decoder, combining the channel attention and spatial attention mechanisms.

In the Encoder part, the network contains multiple 1x1 convolutional layers with 64, 128, 256, 512, and 1024 channels in order, and the convolutional layers are followed by the ReLU activation function and downsampled by a maximum pooling layer (with a step size of 2). Bottleneck is located between the encoder and decoder and contains a 1024-channel 1x1 convolutional layer followed by a ReLU activation function, which is mainly used to provide efficient feature transformation between the encoder and decoder.

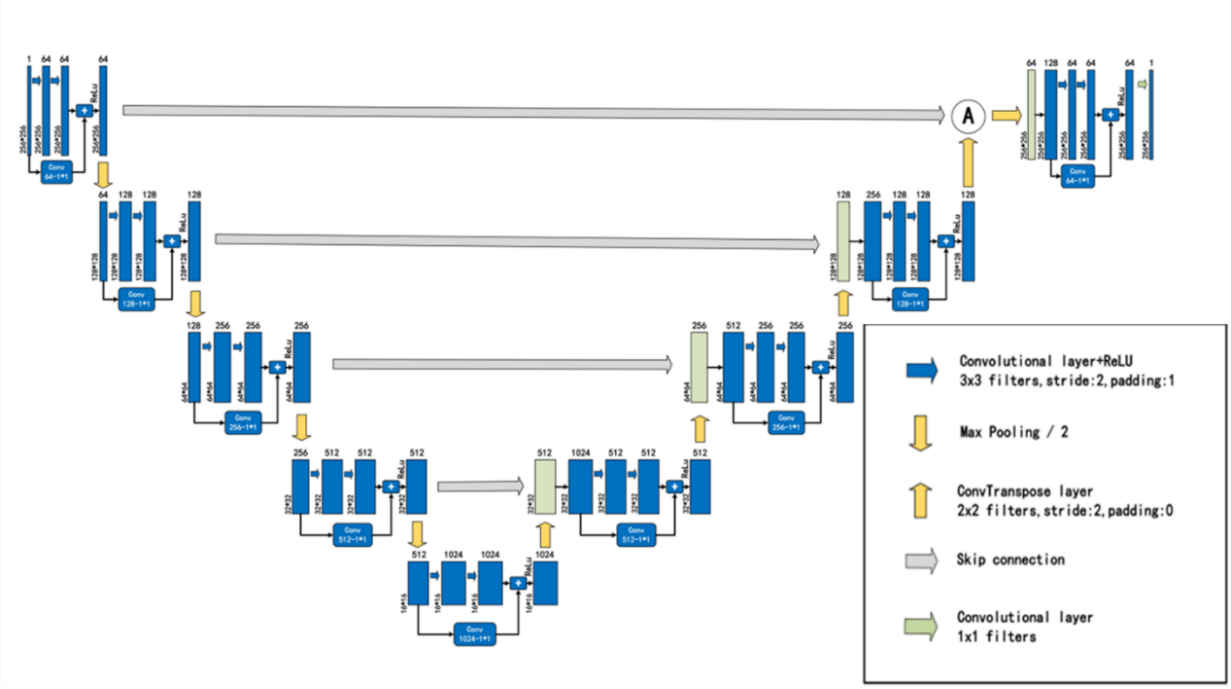

Fig. 1 Channel attention mechanism

The Decoder part then uses a 2x2 convolutional transposition layer for upsampling with a step size of 2 and a padding of 0 to recover the spatial dimensions of the feature map. The Decoder contains multiple 3x3 convolutional layers with channels 512, 256, 128, 64 in that order, and each convolutional layer is followed by a ReLU activation function to increase the nonlinear expressiveness of the network. The encoder and decoder are connected by Skip Connections, which enables the network to better preserve and utilize low-level features, effectively mitigating the gradient vanishing problem and facilitating feature transfer, shown in Fig. 1.

2.3. CSAM Module

Attention mechanisms have been introduced into the network, including channel attention and spatial attention [2,3]. Channel attention adaptively adjusts the weights of each channel, so that the network can better focus on the important feature channels. Spatial attention adaptively adjusts the weight of spatial position, so that the network can better focus on important spatial regions in the image. By adding the attention mechanism, it can more effectively capture and utilize the important features and

details in the image, improve the image segmentation performance, and better process complex background information, improving the accuracy and robustness of segmentation.

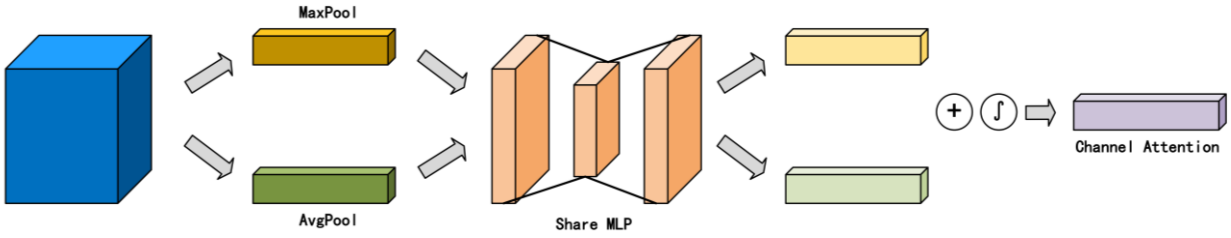

Fig. 2 Channel attention mechanism

The channel attention mechanism mainly focuses on the channels in the input feature map, that is, the different features in the feature map. The basic structure and workflow of adjusting and highlighting important features by assigning weights to each channel is as follows: The input feature plot is as follows, where $X \in R^{C\times H\times W}$ $C, H,$ and W represent the number of channels, height, and width, respectively. Firstly, the Global Average Pooling (GAP) and Global Max Pooling (GMP) were performed on each channel to obtain two eigenvectors, GAP(X) and GMP(X) [4]., their dimensions are: R^C and the two eigenvectors are then fed into a shared Multilayer Perceptron (MLP), usually a neural network containing a hidden layer, to generate channel weights. The output of MLP is a weight vector $W \in R^C$. The final channel attention weight is obtained by additionally fusing the two weight vectors $W_{\mathrm{CA}} = \sigma(W_{\mathrm{avg}} + W_{\mathrm{max}})$, in which σ is the Sigmoid activation function. Finally, the channel weights are multiplied by the original input feature map channel by channel to obtain the weighted feature map $X_{CA} = W_{CA} X$. Through the above process, the channel attention mechanism assigns different weights to each channel, highlights the feature channels that are important to the current task, and improves the representation ability of the network, shown in Fig. 2.

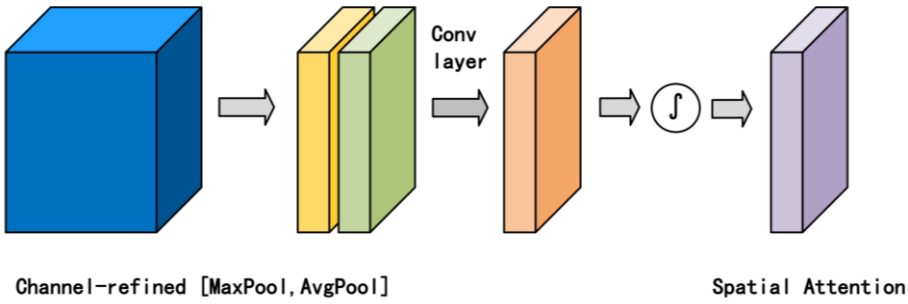

Fig. 3 Spatial attention mechanism

The spatial attention mechanism mainly focuses on the spatial locations in the input feature map, adjusting and highlighting important spatial regions by assigning weights to each spatial location. The basic structure and workflow are as follows: the input feature plots are as follows, where $X \in R^{C\times H\times W}$ $C, H,$ and W represent the number of channels, height, and width, respectively. Firstly, the global average pooling (GAP) and global maximum pooling (GMP) are pooled along the channel dimension [5], and the sum of the two two-dimensional feature maps is obtained $X_{avg} = R^{1\times H\times W}$ and $X_{max} = R^{1\times H\times W}$. Then, the two two-dimensional feature maps are stitched along the channel dimension to obtain a new feature map $X_{concat} = R^{2\times H\times W}$. Through convolution operations, usually using 7×7

convolution kernels, a single-channel attention map is obtained $X_{SA} = R^{1\times H\times W}$. Sigmoid activation was performed on the attention map to obtain spatial weights. Finally, the spatial weights are multiplied by the spatial positions of the original input feature map to obtain the weighted feature map $X_{SA} = W_{SA} X$. Through the above process, the spatial attention mechanism assigns different weights to each spatial position, highlights the important areas in the image, and improves the network's ability to pay attention to key spatial information, shown in Fig. 3.

2.4. Loss Function

In deep learning models, the loss function is the core component of the optimization algorithm, and its main role is to guide the update of model parameters and minimize the difference between the model output and the real label. In this paper, the network model of channel attention mechanism and spatial attention mechanism is combined, and the cross-entropy loss function is selected for the design of the loss function, which is a standard loss function widely used in classification tasks. Specifically, the formula for the cross-entropy loss function is:

$$L(x, y) = -\sum_{i=1}^{c} y_i \log(p_i) \tag{1}$$

Where C represents the number of categories, y_i which is the one-hot encoding of the real label, which is the p_i probability of the ith class predicted by the model, and log is the natural logarithm. For a batch of data, the average of the loss function is used as the final loss calculation:

$$L = \frac{1}{N} - \sum_{n=1}^{N} - \sum_{i=1}^{c} y_{n,i} \log(p_{n,i}) \tag{2}$$

Where N is the number of samples in the batch, y_n is the one-hot code of the true label of the nth sample, and $p_{n,i}$ is the predicted probability of the nth sample to the ith class. The cross-entropy loss function is used to measure the difference between the predicted probability distribution p and the true label distribution y of the model, by taking the logarithm of the prediction probability p_i of each category and multiplying it by the real label y_i , and then summing all classes, the difference between the predicted distribution and the true label distribution is calculated, so as to guide the update of the model parameters.

3. Model Testing

3.1. Datasets

In this study, to evaluate the learning ability and generalization ability of the proposed model, the knee MRI dataset was selected for experimental evaluation. The dataset contains 300 images and their labels. To ensure the reliability of the experiment, we randomly divided the dataset into a training set, a test set, and a validation set with a ratio of 8:1:1, where 80% of the data was used for training, 10% for testing, and 10% for validation. Since the inconsistent resolution of the original images affects the training efficiency, we reduced the resolution of all images to 512×512 pixels in equal proportions and applied smoothing processing and histogram equalization techniques in the

preprocessing stage to reduce the impact of noise on the training effect and improve the robustness and generalization ability of the model [6].

3.2. Experimental Environment

In this experiment, we used the deep learning frameworks PyTorch 2.2.0 and Python 311 and CUDA 12.4 for model training. All experiments were done on servers configured with Xeon(R) Platinum 8474C processors and RTX 4090D graphics cards. To ensure the stability and effectiveness of the training process, the batch size is set to 10, and the training iteration round is set to 500 epochs. The initial learning rate of the model is set to 2e-5, and the optimizer adopts the Adam optimization algorithm. The β parameter setting of the Adam optimizer ranges from 0.9 to 0.999. To improve the training efficiency, the learning rate was adjusted during the training process, i.e. after 300 epochs, the learning rate was reduced by 90%.

3.3. Evaluation indicators

Loss Rate、Mean Absolute Error, MAE、F1 Score are chosen as evaluation indexes to measure the model performance.

The loss rate is an important metric to measure the difference between the predicted results of the model and the actual label. In machine learning and deep learning models, the loss rate is often used as an objective function of the optimization process, improving model performance by minimizing the loss rate. The loss ratio is defined as follows:

$$LossRate = \frac{1}{N}\sum_{i=1}^{N} L(y_i, y_i') \tag{3}$$

Where N represents the total number of samples, y_i represents the true label of the ith sample, represents y_i' the predicted value of the ith sample, and L represents the loss function. The lower the loss rate, the smaller the difference between the model's prediction results and the actual labels.

Mean square error (RMS) is a commonly used evaluation metric in regression problems, which is used to measure the mean absolute error between the predicted value and the actual value. MAE is calculated as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} | y_i - y_i' | \tag{4}$$

Where N represents the total number of samples, y_i represents the true value of the ith sample, and y_i' represents the predicted value of the ith sample. The MAE reflects the average absolute difference between the predicted and true values, with smaller MAE values indicating better predictions.

The $F1$ index is an evaluation metric that takes both precision and recall into account and is especially useful for datasets with unbalanced categories. The formula for calculating the $F1$ index is as follows:

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

Where precision represents the proportion of true positive samples among all samples predicted to be positive:

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

Recall indicates the proportion of samples that are correctly predicted to be positive out of all samples that are truly positive:

$$Precision = \frac{TP}{TP + FN} \tag{7}$$

Among them TP, FP and FN represents True Positives, False Positives, and False Negatives. The F1 index ranges from 0 to 1, with higher values indicating better classification performance of the model. We have compared this with other approaches in the current field. To ensure the accuracy of the experimental results, we used the same training method and dataset settings for all comparison models. As can be seen from the table, our proposed method has shown superior results in the semantic segmentation task of MRI images of knee patellar ligaments, shown in Table 1 and Fig. 4.

Table. 1 Segmentation results of different methods on knee MRI dataset

Method	Loss Rate	THERE ARE	F1 Score
U-Net	0.0400	0.004	0.569
U-Net++	0.0351	0.003	0.636
Res-Net	0.0378	0.003	0.654
FR-UNet	0.0431	0.003	0.624
Ours	0.0321	0.003	0.663

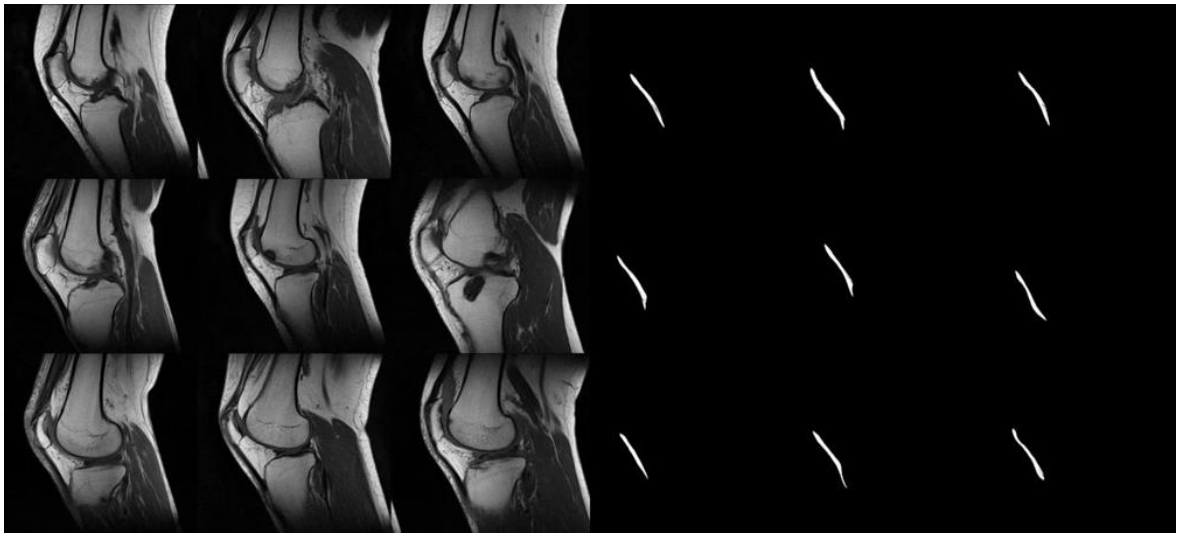

Fig. 4 Results of patellar ligament segmentation

4. Summary

In this paper, an improved U-Net semantic segmentation model based on residual connection and attention mechanism is proposed, and it is comprehensively evaluated and analyzed. The model significantly improves the performance of semantic segmentation tasks by introducing residual

connection, channel attention mechanism and spatial attention mechanism. Compared with the traditional U-Net model, the improved version proposed in this paper has the following innovations: first, it introduces residual connections, which solves the problem of information loss caused by traditional hopping connections, and improves the stability and training effect of the network; Second, the channel and spatial attention mechanism are fused, and the feature response is enhanced by adaptively, and the model's attention to important features is optimized, so as to improve the segmentation accuracy. Accurate segmentation results can be used as an important basis for doctors to judge the condition and improve the reliability and accuracy of diagnosis.

References

[1] Eddie Wei,Qi Luo,Yingzhi Zhao. Semantic segmentation model based on adaptive fusion and attention refinement [J]. Journal of System Simulation,2023,35(06):1226-1234.

[2] Yang Xin,Wang Qiong,YAO Asia,et al. Improved aircraft detection for optical remote sensing images based on Faster R-CNN [J]. Advances in Lasers and Optoelectronics,2023,60(12):427-437.

[3] Huang Jianhua,LI Chaojun,SHA Lei,et al. A plaque segmentation method for 3D carotid ultrasound images by fusing convolutional neural network and Transformer [J]. Mechatronics,2022,28(Z2):71-78.

[4] Yan Haolei,LI Xiaochun,ZHANG Renfei,et al. Pedestrian re-recognition by fusing multiscale attention and bidirectional LSTM [J]. Journal of Air Force Engineering University,2022,23(05):71-76.

[5] Su Xiaodong,LI Shizhou,ZHAO Jiayuan,et al. Image semantic segmentation based on multilevel superposition and attention mechanism [J]. Computer Engineering,2023,49(09):265-271+278.

[6] Zhai Xuming,LI Xiao,ZHAI Yujia. Research on defect detection method of overhead transmission conductor based on deep learning [J]. Grid Technology,2023,47(03):1022-1031.