ISSN: 1813-4890

CSE-TransNet : Cell Nucleus segmentation method with Transformer

Ping Zou, Jiansheng Wu

School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan, 114051, China.

Abstract

To improve the performance of U-Net in cell nucleus segmentation, a CSE-TransNet network is proposed. The CSE-TransNet network structure consists of three parts: encoder, decoder and skip connection. The encoder and de-coder structure uses the UCTransNet network for feature extraction, and a CSE-Transformer method is proposed at the skip connection. This method improves Q and K calculation methods, so that Q can perform self-attention mech-anism calculation on K in the global range, and enhance the ability of the Multi-Head Self-Attention to capture long-distance feature information; In the MLP network, Squeeze-and-Excitation block is added to enhance the chan-nel attention mechanism after the last fully connected layer and deep separable convolution is added to the residual path to extract local neighborhood information. In this way, CSE-Transformer takes into account the context global and the local neighborhood channels information in the feature map. In the MoNuSeg cell nucleus segmentation ex-periment, CSE-TransNet achieved excellent segmentation performance.

Keywords

Cell nucleus segmentation, squeeze-and-excitation, CSE-Transformer, UCTransNet.

1. Introduction

In computer-aided diagnosis and intelligent medicine, in order to improve the efficiency and accuracy of diagnosis, medical segmentation method is often used to show the changes of pathological structure more clearly. Convolutional Neural Network (CNN) is the main method of deep learning when it is just applied in the field of medical segmentation. Fully Convolutional Network (FCN)[1][8], proposed in 2015, has made pioneering achievements in the field of medical segmentation. In recent years, encod-decoder based on convolution has achieved impressive results in image segmentation tasks. Inspired by residual connection and Dense connection, Res-UNet[3] and dense-Unet[4] respectively apply residual connection and dense connection to every submodule of U-Net[2] and alleviate the problem of gradient disappearance in the deep network. On the basis of Res-UNet++[5], UNet++[6] is taken as the baseline model, and extrusion excitation module[22] and attention module are added to make the model pay more attention to the feature information in the region of interest in the feature diagram, thus improving the segmentation performance of the model to a certain extent. Attention-unet[7] replaces hard Attention with soft attention and integrates it into the traditional UNet network to simplify the computational complexity of the model and improve the prediction accuracy of the model.

Using CNN networks has limitations in capturing long-distance dependencies. In recent years, Transformer[9] in a natural language processing task has attracted wide attention in the field of computer vision, and good experimental results have been obtained in tasks such as image classification, object detection[12] and semantic segmentation[10]. The self-attention mechanism in Transformer captures long-distance information by capturing the correlation between tokens. This mechanism can effectively make up for the shortcomings of the network model in obtaining local information from neighborhood pixels through convolution operations and ignoring context-related information. Although CNN network can expand the information of convolution field of view by deepening the number of network layers, Transformer model, even at the lowest layer, can also make the model have a larger field of view through the self-attention mechanism. In 2020, ViT[13] introduced Transformer into the computer vision task for the first time. Different from CNN or ResNet's convolution core of fixed size, the initial layer of ViT has a larger vision. Many scholars have introduced Transformer and its variants[11] into visual tasks across domains, bringing creative effects and a significant increase in accuracy. Tolstikhin[21] et al. proposed the MLP-mixer model in 2021, which used the Multi-Layer Perceptron (MLP) network to mix the information among image blocks, and combined the mixed image information through the superposition of these information blocks. It can be seen that improvements to MLP will greatly improve the network performance of Transformer. Petit[17] et al. used the improved multi-head self-attention mechanism at the bottom of U-Net network and the skip connection to enhance the global information interaction between network layers and to recover and filter out non-semantic features. The improved multi-head self-attention mechanism greatly improved the segmentation performance.

Although Transformer[26] has achieved good results in medical segmentation tasks, it still has limitations in some complex tasks. This paper proposes an efficient Transformer method and applies it to UCTransNet[24] network, which is called CSE-TransNet model. The network not only enhances the ability to obtain the long distance information in the feature map, but also can take into account the local channel information around the feature. The network has achieved good performance in the MoNuSeg nuclear medical image segmentation task.

2. Method

2.1. CSE-TransNet

The CSE-TransNet structure is shown in Fig 1. The encoder structure of this model is consistent with that of U-Net network. The input feature $\operatorname{Im} g \in R^{3\times H\times W}$ gets the coding layer feature matrix $E_i \in R^{C_i\times\frac{H}{2^{i-1}}\times\frac{W}{2^{i-1}}}$，（i=1,2,3,4,5）through continuous convolution and downsampling. Within the first four layers, 2D patch sequences $T_i \in R^{\frac{HW}{p^2}\times C_i}$ are obtained by Embeded using patches of different sizes (p= 16,8,4,2). Splice these four layers of T_i sequences to get $T_\Sigma = (T_1, T_2, T_3, T_4)$ into CSE-Transformer, T_i as the query of the multi-head self-attention mechanism, T_Σ as the key and value,the characteristic tensor $O_i \in R^{C\times H\times W}$ is obtained by self-attention calculation. In order to effectively connect to the decoding layer to eliminate feature ambiguity, O_i and decoding layer D_i are combined to obtain O_i through Channel-wise Cross Attention module[24] (CCA). The feature information of O_i and D_i is fused in the decoding layer, and the final segmentation figure $SegMap \in R^{H\times W}$ is obtained through 1×1 convolution and Sigmoid function.

2.2. CSE-Transformer

CSE-Transformer used in jump connections is mainly composed of CSE-MHSA (CSEMulti-Head Self-Attention), CSEMulti-Layer Perceptron, CSEMulti-MLP, CSE-MLP) and Layer Normalization (LN). The layer normalization operation was used before the CSE-MHSA and CSE-MLP modules, and the output results of these two modules were added to the feature maps before the normalization operation using the residual join. Fig 2 shows the CSE-Transformer network.

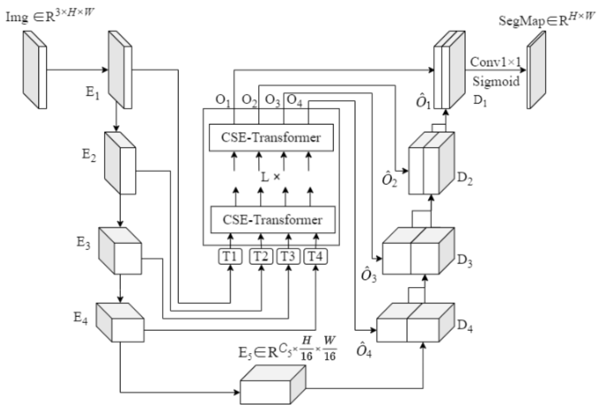

Fig.1 CSE-TransNet structure

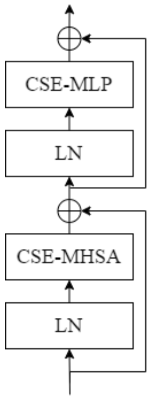

Fig.2 The overall framework of CSE-Transformer

The calculation formula is as follows:

$$y' = x + CSE - MHSA\big(LN(x)\big) \tag{1}$$

$$y = y' + CSE - MLP\big(LN(y')\big) \tag{2}$$

2.2.1. CSE-MHSA

In the original Transformer, multiple self-attention operations[16] enable the model to obtain rich feature information and provide multiple representation subspaces to learn the weight matrix of query vectors query, key and value, and dot multiply Q with K in the whole sequence to obtain self-attention scores. Modeling the correlation between all tokens in the self-attention mechanism and using the information between channels to search for information between global space and channels enables the model to focus on medium and long distance information in the feature graph information. Finally, add the SoftMax function to prevent the disappearance of gradient, and multiply with the V matrix to get the final output structure. The calculation formula is as follows:

$$Attention(Q,K,V) = SoftMax\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{3}$$

Although Transformer shows great research value compared with CNN network, it still has some shortcomings: direct dot product of Q and K obtained after flattening in multi-head self-attention mechanism may reduce the correlation between feature channels; Each header is only responsible for inputting a subset of tokens. This operation may affect the network performance. Especially, when the channel dimension of each subset is very low, query and key dot products can no longer match information, thus reducing the ability to obtain context information. For some high resolution images, the amount of computation increases with the size of the space or channel. To solve these problems, we propose a CSE-Transformer multiple self-attention mechanism. By improving the feature matrix obtained after the different input query and key in each header and the dot product, the method is flexible, without interpolation or fine tuning, and the self-attention computation efficiency is improved while maintaining the diversity ability of multiple self-attention mechanisms. The CSE-MHSA network is shown in Fig 3.

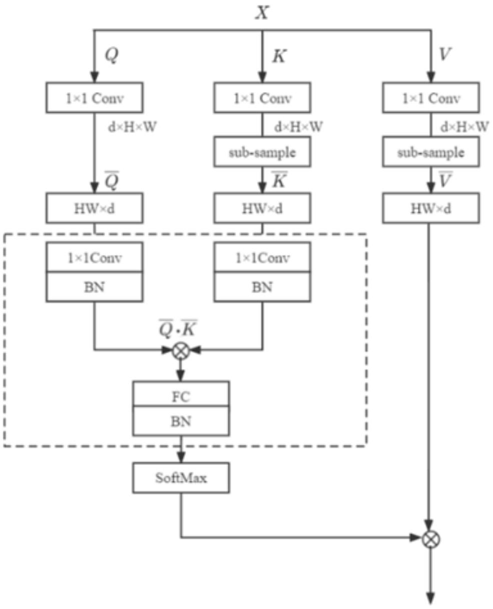

Fig.3 CSE-MHSA network

Input the feature tensor $X \in R^{H\times W\times C}$, where H and W are the height and width respectively, and C is the number of channels. You take X and you projection it through three 1 by 1 convolution to get $Q,K,V \in R^{d\times H\times W}$, *d* is *Q*,and the dimension of the *K* vector. *Q*, *K* and *V* are transformed into sequence length of is after flattening, and 1×1 convolution is added to enhance the interaction between *Q* and *K* in multiple autoattention to obtain local context information. The relationship between *Q* and K is not only modeled separately, but also cross-channel feature aggregation is carried out. After the dot product of *Q* and *K*, the use of full joins to expand the channel dimension and prevent the reduction of dimension enables the results after the dot product to match the feature information in the vector *V*. Although this operation compromises the ability of the self-attention mechanism to combine processing information from different subsets in different locations, it is followed by Batch Normalization. BN) to restore this capacity for diversity. The new attentional mechanism Atten (*Q*,*K*,*V*) can be calculated as follows:

$$Atten = S\left(B\left(FC\left(\frac{\left[B\left(C(Q)\right)\right]\otimes\left[B\left(C\left(K^{T}\right)\right)\right]}{\sqrt{d_k}}\right)\right)\right)V \tag{4}$$

S stands for SoftMax function, *B* for batch normalization operation, FC for full connection, and C for 1×1 convolution.

2.2.2. CSE-MLP

On the basis of retaining the MLP network structure, the CSE-MLP network enhances the correlation between local channels by adding the extrusion excitation module SE block and depth separable convolution 0. The CSE-MLP network is shown in Fig 4. After fully connected FC2, the SE block module is added, which includes extrusion and excitation operation. The extrusion operation is to integrate the feature matrix output by the linear layer in space dimension and enhance the global feature information of the network. Excitation operation is to enhance the degree of dependence between channels, and according to the recent research work[23], depth separable convolution and linear layer FC3 are added to the residual path to capture local information and enhance the CSE-MLP network's ability to obtain local feature information.

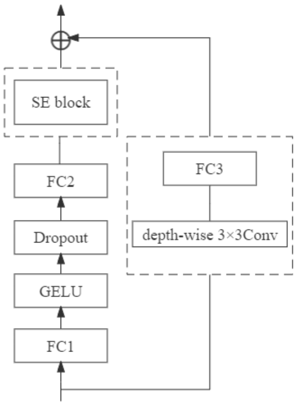

Fig.4 CSE-MLP Network

The CSE-MLP network calculation formula is as follows:

$$MLP(x) = SE\left(FC2\left(D\left(G\left(FC1(x)\right)\right)\right)\right) + FC3\left(DW\left(x\right)\right) \tag{5}$$

D represents the Dropout operation, G represents the GELU activation function, and DW represents the 3×3 depth separable convolution.

3. Experiment

3.1. Experimental environment

The experimental operating system in this paper is Ubuntu18.4, the GPU model is NVIDIA 3060, and the video memory size is 12GB. The network framework is pytorch 1.8, and the programming languages are python 3.8 and CUDA 11.1. The MoNuSeg[25] dataset was used to create a dataset for nuclear segmentation based on H&E stained tissue images captured at 40x magnification. The training set contains 30 images and training data of about 22,000 nuclear boundary annotations, and the test set contains 14 images and test set images with an additional 7000 nuclear boundary annotations.

3.2. Evaluating metric

The evaluation index adopted in this paper is the Dice Coefficient[14] (Dice Similarity Coefficient (DSC) and the IOU index (Intersection Over Union (IOU), which is commonly used in medical image segmentation. The calculation formula is as follows:

$$Dice = \frac{2|X \cap Y|}{|X| + |Y|} \tag{6}$$

$$IOU = \frac{|X \cap Y|}{|X \cup Y|} \tag{7}$$

Where X and Y represent the pixel matrix in the real and predicted images respectively, Dice represents the overlapping part in the predicted and predicted images, and is used to calculate the similarity. IOU is used to calculate the ratio of the intersection and union of two sets of true and predicted values. The value of the evaluation indicator ranges from 0 to 1. A larger value indicates better network performance.

3.3. Experimental results and analysis

In this paper, two types of methods are compared, one is the combination of CNN U-Net network, such as U-Net, UNet++, AttentionUNet, MRUNet; The other category is U-Net networks combined with Transformer, such as TransUNet, MedT, Swin-Unet and UCTransNet. The input image resolution was set as 224×224, the Adam optimizer was adopted, the initial learning rate was 1e-3, the batch size was 4, and the network model was trained by combining the cross entropy loss function and Dice loss function. The experimental results are shown in Table 1. Compared with the latest UCTransNet model method, the evaluation index Dice and IOU of CSE-TransNet proposed in this paper increased by 1.33% and 1.85%.

Table 1 Experimental results

Method	Dice(%)	IOU(%)
U-Net	76.45	62.86
UNet++	77.01	63.04
Attention-UNet	76.67	63.47
MRUNet	78.22	64.83
TransUNet	78.53	65.05
MedT	77.46	63.37
Swin-Unet	77.69	63.77
UCTransNet	79.08	65.50
CSE-TransNet	80.41	67.35

We visualized the segmentation result of the model, as shown in Fig 5. (a) represents the real label, (b) represents the segmentation result graph of UCTransNet model, and (c) represents the CSE-TransNet segmentation result graph. Compared with the latest UCTransNet model, CSE-TransNet is closer to the real label, showing good segmentation results.

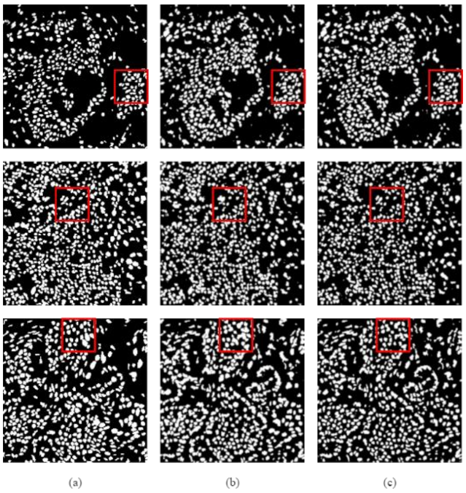

Fig.5 The corresponding quantitative results

4. Conclusion

Medical image segmentation plays an important role in computer vision tasks. This paper proposes a CSE-TransNet structure for nuclear image segmentation. The UCTransNet model method is adopted for the backbone network, but the CSE-Transformer structure is designed at the network jump connection. On the one hand, the multi-head self-attention mechanism is strengthened to search the features in the global range for the long distance range. On the other hand, the MLP structure can segment some local features in the neighborhood more accurately by introducing the extrusion excitation module and adding the depth separable convolution layer and the full connection layer to the residual path. It was tested on the MoNuSeg dataset, and the experimental results show that the proposed algorithm has high accuracy and can segment the nuclear lesion region well. In future studies, other segmentation tasks such as liver tumor segmentation, lung nodules[15] and brain tumors[19] will be used to verify whether the model proposed in this paper is also applicable. In the follow-up research, how to further improve the accuracy of segmentation, optimize the network model and adjust the number of network parameters is also the focus of the future research direction.

References

[1] Long J, Shelhamer E and Darrell T: Fully Convolutional Networks for Semantic Segmentation. IEEE Transac-tions on Pattern Analysis and Machine Intelligence, Vol. 39 (2015) No.4,p.640-651.

[2] Ronneberger O, Fischer P and Brox T: U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention. (Munich, Germany, October, 2015). P. 234.

[3] Ibtehaz N, Rahman M S: MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image seg-mentation. Neural Networks, Vol. 121 (2020), p. 74-87.

[4] Li X, Chen H, Qi X, et al: H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. IEEE transactions on medical imaging, Vol.37 (2018) No. 12, p.2663-2674.

[5] Jha D, Smedsrud P H, Riegler M A, et al: Resunet++: An advanced architecture for medical image segmenta-tion. 2019 IEEE International Symposium on Multime-dia (ISM). (ShangHai, China, July 8-12, 2019). Vol. 1, p.225.

[6] Zhou Z , Siddiquee M M R , Tajbakhsh N , et al: UNet++: A Nested U-Net Architecture for Medical Image Segmentation. 4th Deep Learning in Medical Image Analysis (DLMIA) Workshop. (Spain,September 20,2018).p.223.

[7] Oktay O , Schlemper J , Folgoc L L , et al: Attention U-Net: Learning where to look for the pancreas .MIDL, (Amsterdam, Netherlands, July 4-6, 2018). Vol. 1, p.1.

[8] Milletari F, Navab N, Ahmadi S A: V-net: Fully convolu-tional neural networks for volumetric medical image segmentation. 2016 fourth international conference on 3D vision (3DV). (Stanford, USA, October 25-28, 2016). p.565.

[9] Vaswani A, Shazeer N, Parmar N, et al: Attention is all you need. Advances in neural information processing systems. (Long Beach, USA, December 3-9, 2017). p.5998.

[10] Valanarasu J M J, Oza P, Hacihaliloglu I, et al: Medical transformer: Gated axial-attention for medical image seg-mentation. International Conference on Medical Image Computing and Computer-Assisted Intervention. (Strasbourg, France, September 27-October 1, 2021). p.36.

[11] Liu Z, Lin Y, Cao Y, et al: Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF International Conference on Computer Vision. (ChangSha, China, August 20-22,2021). p.10012.

[12] Carion N, Massa F, Synnaeve G, et al: End-to-end object detection with transformers. European Conference on Computer Vision. (Glasgow, UK, August 23-28, 2020). p.213.

[13] Yuan L, Chen Y, Wang T, et al: Tokens-to-token vit: Train-ing vision transformers from scratch on imagenet. Proceedings of the IEEE/CVF International Conference on Computer Vision. (ChangSha, China, August 20-22,2021). p. 558.

[14] Dice L R: Measures of the amount of ecologic association between species. Vol. 26 (1945) No. 3, p. 297-302.

[15] Kubota T, Jerebko A K, Dewan M, et al: Segmentation of pulmonary nodules of various densities with morphological approaches and convexity models. Vol. 15 (2011) No. 1, p. 133-154.

[16] Zhao H, Jia J, Koltun V: Exploring self-attention for image recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (XiAn, China, August 7-9, 2020). p.10076.

[17] Petit O, Thome N, Rambour C, et al: U-net transformer: Self and cross attention for medical image segmenta-tion. International Workshop on Machine Learning in Medical Imaging. (Strasbourg, France, September 27-27, 2021).p.267.

[18] Tang Z, Jiang W, Zhang Z, et al: DenseNet with Up-Sampling block for recognizing texts in images. Neural Computing and Applications. Vol. 32 (2020) No. 11, p. 7553-7561.

[19] Zhang H L, Li Q, Guan X: An improved three dimensional dual-path brain tumor image segmentation network. Vol. 41 (2021) No. 3, p. 0310002.

[20] Zeiler M D, Taylor G W, Fergus R: Adaptive deconvolutional networks for mid and high level feature learning. 2011 international conference on computer vision. (Barcelona, Spain, November 6-13, 2011). p. 2018.

[21] Tolstikhin I O, Houlsby N, Kolesnikov A, et al: Mlp-mixer: An all-mlp architecture for vision. Advances in Neural Information Processing Systems. Vol. 34 (2021), p. 24261-24272.

[22] Hu J, Shen L, Sun G: Squeeze-and-excitation net-works. Proceedings of the IEEE conference on computer vision and pattern recognition. (Salt Lake City, USA, June 18-21, 2018).p. 7132.

[23] Wang Z, Cun X, Bao J, et al: Uformer: A general u-shaped transformer for image restoration. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.(New Orleans, Louisiana, June 19-23, 2022). p. 17683.

[24] Wang H, Cao P, Wang J, et al: Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. Proceedings of the AAAI Conference on Artificial Intelligence. (Vancouver, Canada, February 22- March 1, 2022). Vol. 36. p.2441.

[25] Kumar N, Verma R, Sharma S, Bhargava S, Vahadane A, Sethi A: A Dataset and a Technique for Generalized Nu-clear Segmentation for Computational Pathology. IEEE Transactions on Medical Imaging. (Nice, France, April 13-16, 2017). Vol.36. p. 1550.

[26] Liu Wen-Ting, Lu Xing-Ming: Research Progress of Transformer Based on Computer Vision. Computer Engineering and Applications. Vol.58 (2022) No.6, p.1-16.