FEATURE SPACE MESSAGE PASSING NETWORK FOR MEDICAL IMAGE SEMANTIC SEGMENTATION

Junxiao Sun[1], *Ke Zhang*[1], *Shuyi Niu*[1], *Yan Zhang*[1], *Youyong Kong*[1,2,⋆]

[1]Jiangsu Provincial Joint International Research Laboratory of Medical Information Processing, School of Computer Science and Engineering, Southeast University, China
[2]Key Laboratory of Computer Network and Information Integration, Southeast University, Ministry of Education, Nanjing, China
⋆Corresponding Author. E-mail:kongyouyong@seu.edu.cn

ABSTRACT

Accurate semantic segmentation of medical images is of significant importance for subsequent processing and analysis. The encoder-decoder deep learning framework has been widely applied for numerous medical image segmentation tasks. However, most existing approaches are restricted by the limited receptive field for failing to capture long-range dependencies, meanwhile lacking global features for spatial information recovery. To solve both problems, we propose a novel feature space message passing network (FSMPN) framework. At first, a dynamic message passing block (DMPB) is proposed to perform the long-range interactions for better feature learning between voxels. Secondly, a skipped graph connection (SGC) module is developed to explicitly transfer learned graph with features from encoder stage to decoder stage to help recover spatial information. The proposed FSMPN was able to achieve superior performance on different types of medical image datasets compared to other popular models.

Index Terms— Medical Image, Semantic Segmentation, Message Passing, Graph Convolution, Long-range Information

1. INTRODUCTION

Semantic segmentation of medical images is critical for further processing and analysis. The encoder-decoder deep learning framework has been widely applied to medical image semantic segmentation. Fully convolutional network establishes the encoder-decoder framework with skip connection for image segmentation [1, 2]. The encoder part extracts features with convolution and pooling operators, and the decoder recovers the spatial information with up-sampling operators. After that, a number of works have been proposed to improve the performance of the encoder-decoder framework in two ways.

The first way is to enhance the feature extraction in the encoder part with introducing effective convolution or pooling modules [3, 4]. Most of the existing approaches utilize convolution kernels with limited size. The CNN framework aggregates long-range information by stacking convolutional layers to make the far apart voxels in a receptive field. However, the limited receptive field cannot well capture long-range dependencies. The medical images always have similar voxels with long distance. It is necessary to explore the relationships between these voxels. To overcome this limitation, the Non-local network [5] designed a self-attention block to capture such long-range relations in the whole feature space. Fortunately, the graph based message passing has a great advantage in capturing long-range dependencies due to the non-Euclidean geometric structure [6].

The second way is to improve the effect of spatial information recovery. SegNet [7] records the pooling position by introducing a maximum pooling index. Feature maps of FCN at different scales in the decoder path are summed with down-sampling feature maps [2] while U-Net concatenates them [1]. The residual structure [8] can be seen as a short skip connection allowing parameters to be updated deeply in the network. However, all these models do not make full use of the global information to improve the performance.

To tackle the aforementioned problems, we propose a novel feature space message passing network (FSMPN) for medical image semantic segmentation. Firstly, a dynamic message passing block (DMPB) is proposed to perform the long-range interactions. Secondly, a skipped graph connection (SGC) module is proposed to explicitly transfer learned graph with features from the encoder stage to the decoder stage. Extensive experiments on two medical image segmentation tasks show outstanding performance improvement for the medical semantic segmentation task compared with other excellent methods.

This work is supported by National Key Research and Development Program of China (No. 2021ZD0113202) and grant 31800825 National Natural Science Foundation of China.

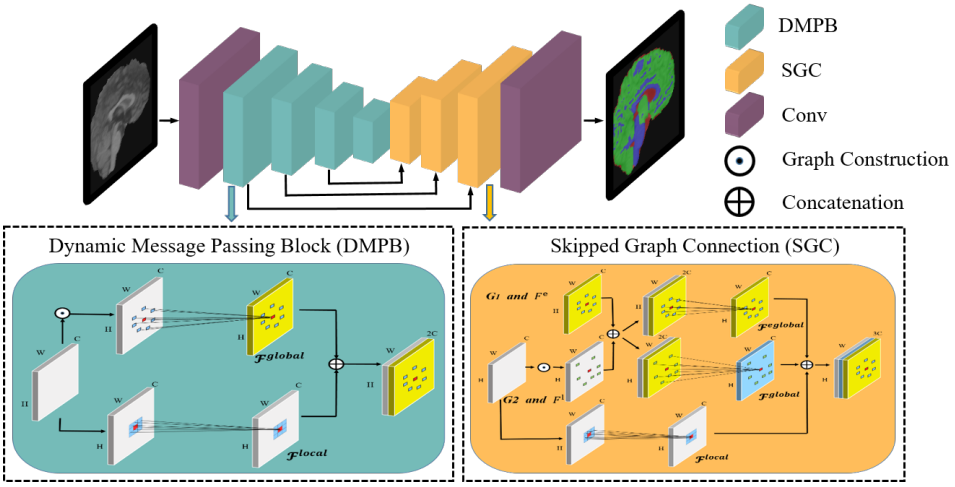

Fig. 1. Illustration of our proposed FSMPN for medical images semantic segmentation. In DMPB, graph convolution acts in the feature space directly to capture long-range relations between voxels. SGC module is employed to lessen the long-range information loss in the decoder path.

2. METHOD

Fig.1 illustrates the overview of the proposed FSMPN segmentation framework which consists of the DMPB and SGC.

2.1. Dynamic Message Passing Block

The bottom left part of Fig.1 shows the architecture of the proposed DMPB which consists of three parts. Firstly, a dynamic graph is constructed to describe the long-range relationships between voxels. Secondly, message passing is performed on the voxel-wise graph to achieve the long-range interactions between voxels. Finally, the features obtained from local convolutions and global long-range interactions are concatenated together.

With the feature maps at each stage, a dynamic voxel-wise graph is constructed to represent the long-range relationships between voxels. The graph is constructed from the feature maps $X^l \in \mathbb{R}^{H\times W\times C}$, where H, W and C denote the height, weight and number of channels of the feature map. The graph at the l-th layer can be represented as $G^l = (V^l, F^l, A^l)$. The nodes V^l represent the voxels, and the number of nodes is $N = H \times W$. F^l represents node features, and C indicates the dimension of F^l. The Euclidean distance [9] is utilized to compute the similarity between each node. The top K nearest neighbors of each node is selected to construct the adjacency matrix A^l as follow:

$$A^l_{i,j} = \begin{cases} 1, & \text{if } j \text{ is one of } K \text{ nearest neighbors of } i \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

Meanwhile, the adjacency matrix A^l constructed in this way is not a symmetric matrix, so G^l is the directed graph which can both guide the direction of message passing in the feature space and balance the number of the neighbors of each node. Since the feature maps are different between each layer, the constructed graphs are dynamic.

Based on the constructed voxel-wise graph, message passing is performed to make the long-range interactions. The message passing operator is defined as:

$$F^{l+1}_i = \sigma(F^l_i W^l_i + \sum \frac{A_{i,j} F^l_j W^l_j}{K^l_j} + b^l) \quad (2)$$

where F^l_i denotes the feature vector of node i in the l-th layer. $W^l_i \in \mathbb{R}^{C\times d^{l+1}}$, $W^l_j \in \mathbb{R}^{C\times d^{l+1}}$ and $b^l \in \mathbb{R}^{1\times d^{l+1}}$ are matrix transformation parameters, which are shared for each node in the l-th layer. d^{l+1} is the dimension of F^{l+1} and K^l_i is the number of neighbors of node i. σ is the $ReLU$ non-linearity function.

In particular, through several convolutional units and dynamic message passing blocks in the l-th layer, we obtain the local features $\mathcal{F}^{local}$ and the global features $\mathcal{F}^{global}$. Then, we combine the local information and long-range dependencies by concatenating the local and global features as follow:

$$\mathcal{F}^{l+1} = \phi(\mathcal{C}(\mathcal{F}^{local}, \mathcal{F}^{global})) \quad (3)$$

where $\phi(\cdot)$ is a 1×1 convolution and $\mathcal{C}(\cdot,\cdot)$ is the concatenation operator. The purpose of 1×1 convolution is not only to achieve a linear combination of features of two branches, but also to reduce the dimension of feature maps in the next layer. The refined feature map $\mathcal{F}^{l+1}$ is the output of this layer.

2.2. Skipped Graph Connection

To transfer the global information for improving spatial information recovery in decoder stage, we design a novel SGC

module. The SGC module explicitly passes the concatenated features and the graph structure in the encoder layer to the corresponding decoder layer to transmit spatial long-range information.

As show in the bottom right part of Fig.1 , there are three branches in the up-sampling stage. Compared with the encoder path, in the decoder stage, a special branch is introduced to compensate for the defect of the up-sampling operation. To lessen the long-range information loss in the decoder path, the graph structure and the features of the nodes are obtained from the corresponding layer in the down-sampling stage of the current layer. In particular, the graph nodes features of the corresponding layer are concatenated with the features in the previous layer to provide a shortcut for gradient flow in the message passing process [10]. The message passing operator in the decoder path is defined as:

$$F_i^{l+1} = \sigma(\mathcal{C}(F_i^l, F_i^e)W_i^l + \sum \frac{A_{i,j}\mathcal{C}(F_j^l, F_j^e)W_j^l}{K_j^l} + b^l) \tag{4}$$

where $W_i^l \in \mathbb{R}^{2C\times d^{l+1}}$, $W_j^l \in \mathbb{R}^{2C\times d^{l+1}}$ and $b^l \in \mathbb{R}^{1\times d^{l+1}}$ are the matrix transformation parameters, which are shared for each node in the l-th layer. F^e is the concatenated feature obtained from the corresponding blocks of encoder layer.

There are two graph structures, G_1 and G_2, where G_1 represents the corresponding node feature relationships in the encoder path, and G_2 is the information about the nearest neighbors of the node in the current layer. We introduce message passing on two branches separately. Specifically, the graph construction for the current layer employs the concatenated feature maps. The purpose is to introduce the long-range information of the corresponding layer and the previous layer into the judgment of the nearest neighbors at the same time, and strengthen the confidence of the nodes of message passing. The graph structure from the encoder path provides the relationships with long-range information between nodes in the embedding process. Similarly, the convolution operations are used to extract local features. Then, after cascading the output of the three branches, a linear combination is achieved by 1×1 convolution as:

$$\mathcal{F}^{l+1} = \phi(\mathcal{C}(\mathcal{F}^{local}, \mathcal{F}^{global}, \mathcal{F}^{eglobal})) \tag{5}$$

where $\mathcal{F}^{eglobal}$ is obtained after the DMPB with the concatenated features and the topological structure of the graph acquired from the corresponding blocks of encoder layer.

3. EXPERIMENTS

In this section, we firstly introduce datasets, implementation details of the proposed FSMPN and the evaluation methods. Then the results on two medical images semantic segmentation tasks of brain tissue segmentation [11] and tumor segmentation [12] are reported. The proposed approach is compared with a series of methods for evaluation. Moreover, we conduct ablation studies for evaluating the modules.

3.1. Datasets

IBSR18. The IBSR18 dataset consists of 18 real magnetic resonance imaging (MRI) volumes. Each data consists of $256 \times 128 \times 256$ voxels [11, 13]. The dataset provides the ground truth for the segmentation of cerebro spinal fluid (CSF), grey matter (GM) and white matter (WM).

BraTS2015. The BraTS2015 dataset contains 274 subjects. The size of each MRI image is $155 \times 240 \times 240$. The evaluation system separates the tumor structure into three regions due to practical clinical applications [12, 14]: (i)whole tumor (WT)(ii)tumor core (TC)(iii)enhancing tumor (ET).

3.2. Implement Details and Evaluation Metric

For all experiments, we trained 60 epochs and selected the epoch with the best performance as the final result. Besides, min-max normalization is selected to linearly scale the image and map the data to the range 0 to 1 to resolve the possible differences in voxel values between different subjects. The Adam optimizer is utilized in the training. The standard cross entropy loss is employed, and the learning rate is set to 1×10^{-5}. The batch normalization is adopted in all experiments, and the learnable weights in our model are initialized with a normalized initialization strategy proposed by [15].

The results are evaluated against the ground truth with the dice similarity coefficient (DSC) which is commonly used to determine the performance of image segmentation. DSC is calculated as:

$$DSC = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{6}$$

where TP, FP and FN are true positive, false positive and false negative voxels. The higher value of DSC indicates a better segmentation result.

3.3. Results

We compare FSMPN with a series of semantic segmentation baselines: U-Net [1], FC-DenseNet [16, 20], ENet [17] and DeepLabV3 [18]. Futhermore, to validate the effectiveness of the proposed long-range information extraction mechanism, we also compare our method with two studies proposed to improve the long-range modeling ability by global message passing(Non-local U-Net) [5, 19] and spatial pyramid graph reasoning(SpyGR) [6]. Table 1 demonstrates our test set results on two medical images datasets. It can be seen that the proposed model shows superior performance to other CNN based and long-range modeling methods for the tumor segmentation on the BraTS2015 dataset. For the brain tissue segmentation task on IBSR18 dataset, it performs well in the segmentation of CSF and GM, while ENet has the better

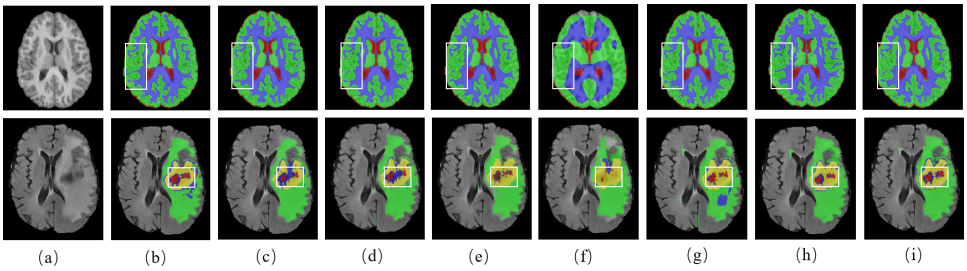

Fig. 2. Visualization of segmentation results for different methods on IBSR18(first row) and BraTS2015(second row). (a) original image, (b) ground truth segmentation, (c) U-Net, (d) FC-DenseNet, (e) ENet, (f) DeepLabV3, (g) Non-local U-Net, (h) SpyGR, (i) the proposed FSMPN.

Table 1. Performance of different segmentation methods on IBSR18 and BraTS2015 datasets measured by DSC(%).

Methods	IBSR18			BraTS2015		
	CSF	GM	WM	WT	TC	ET
U-Net[1]	78.46	93.09	89.64	85.98	66.69	62.97
FC-DenseNet[16]	77.85	92.15	86.35	85.86	67.33	66.72
ENet[17]	76.92	93.07	**90.18**	85.19	55.09	57.63
DeepLabV3[18]	34.07	81.14	69.28	82.42	59.79	50.64
Non-local U-Net[19]	78.14	93.15	89.59	84.21	68.37	65.75
SpyGR[6]	70.21	90.43	86.49	86.45	67.54	66.90
FSMPN	**79.19**	**93.30**	89.72	**86.89**	**71.59**	**69.03**

DSC in WM. Qualitative evaluations of the proposed FSMPN architecture with other comparative models are further assessed, and the visualization results are illustrated as Fig.2, which indicate long-range information enables the segmentation region of the model to be

3.4. Ablation Studies

Ablation studies are further conducted on the BraTS2015 dataset to explore the contribution of each part in FSMPN. At first, we change the architecture of the network and introduce the DMPB to the feature space of different depths on the benchmark U-Net. FSMPN-1l means that the DMPB is employed to the valley of benchmark. By analogy, other structures are shown in Fig.3. The comparisons of DSC value are listed in Table 2 which shows that the a larger number of dynamic message passing blocks could obtain better segmentation performance. Furthermore, we design experiments to verify the effectiveness of the SGC module. As stated above, FSMPN-7l has the best performance. Consequently, we perform ablation experiment on FSMPN-7l. The SGC module is removed from the FSMPN-7l to test the validity of the SGC module. The last two rows of Table 2 demonstrate that the performance has a further increment with the SGC module. more precise.

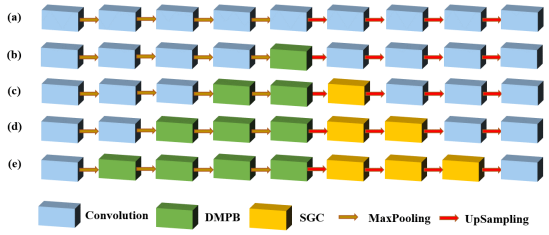

Fig. 3. The architectures of the different depth of the feature space where the message passes. Row (a-e) represent the benchmark, FSMPN-1l, FSMPN-3l, FSMPN-5l and FSMPN-7l, respectively.

Table 2. The DSC(%) of different architectures.

Architecture	BraTS2015		
	WT	**TC**	**ET**
benchmark	85.98	66.69	62.97
FSMPN-1l	86.17	65.61	63.61
FSMPN-3l	86.64	67.71	65.55
FSMPN-5l	**86.94**	71.14	67.07
FSMPN-7l	86.89	**71.59**	**69.03**
w/o SGC	86.59	69.55	68.36

4. CONCLUSION

In this paper, we propose FSMPN, a robust segmentation framework for medical image semantic segmentation. The DMPB is proposed to directly perform the long-range interactions. The SGC module is developed to pass global message between corresponding blocks of encoder and decoder layers. Compared with CNN based and long-range modeling methods, FSMPN successfully achieves precise segmentation results on two medical image segmentation tasks.

5. REFERENCES

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[2] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[3] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[4] Anurag Arnab, Sadeep Jayasumana, Shuai Zheng, and Philip HS Torr, "Higher order conditional random fields in deep neural networks," in *European Conference on Computer Vision*. Springer, 2016, pp. 524–540.

[5] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[6] Xia Li, Yibo Yang, Qijie Zhao, Tiancheng Shen, Zhouchen Lin, and Hong Liu, "Spatial pyramid based graph reasoning for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8950–8959.

[7] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[9] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[10] Michal Drozdzal, Eugene Vorontsov, Gabriel Chartrand, Samuel Kadoury, and Chris Pal, "The importance of skip connections in biomedical image segmentation," in *Deep learning and data labeling for medical applications*, pp. 179–187. Springer, 2016.

[11] Yan Zhang, Youyong Kong, Jiasong Wu, Coatrieux Gouenou, and Huazhong Shu, "Brain tissue segmentation based on graph convolutional networks," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 1470–1474.

[12] Kuan-Lun Tseng, Yen-Liang Lin, Winston Hsu, and Chung-Yang Huang, "Joint sequence learning and cross-modality convolution for 3d biomedical segmentation," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 6393–6400.

[13] Junxiao Sun, Yan Zhang, Jian Zhu, Jiasong Wu, and Youyong Kong, "Semi-supervised medical image semantic segmentation with multi-scale graph cut loss," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 624–628.

[14] Fan Xu, Haoyu Ma, Junxiao Sun, Rui Wu, Xu Liu, and Youyong Kong, "Lstm multi-modal unet for brain tumor segmentation," in *2019 IEEE 4th International Conference on Image, Vision and Computing (ICIVC)*. IEEE, 2019, pp. 236–240.

[15] Xavier Glorot and Yoshua Bengio, "Understanding the difficulty of training deep feedforward neural networks," *Journal of Machine Learning Research*, vol. 9, pp. 249–256, 2010.

[16] Simon Jégou, Michal Drozdzal, David Vazquez, Adriana Romero, and Yoshua Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 11–19.

[17] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," *arXiv preprint arXiv:1606.02147*, 2016.

[18] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[19] Zhengyang Wang, Na Zou, Dinggang Shen, and Shuiwang Ji, "Non-local u-nets for biomedical image segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 6315–6322.

[20] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.