Automatic Segmentation of Brain Tumor Using 3D SE-Inception Networks with Residual Connections

Hongdou Yao, Xiaobing Zhou(✉), and Xuejie Zhang

School of Information Science and Engineering, Yunnan University,
Kunming 650091, People's Republic of China
zhouxb@ynu.edu.cn

Abstract. Nowadays, there are various kinds of methods in medical image segmentation tasks, in which Cascaded FCN is an effective one. The idea of this method is to convert multiple classification tasks into a sequence of two categorization tasks, according to a series of sub-hierarchy regions of multi-modal Magnetic Resonance Images. We propose a model based on this idea, by combining the mainstream deep learning models for two dimensional images and modifying the 2D model to adapt to 3D medical image data set. Our model uses the Inception model, 3D Squeeze and Excitation structures, and dilated convolution filters, which are well known in 2D image segmentation tasks. When segmenting the whole tumor, we set the bounding box of the result, which is used to segment tumor core, and the bounding box of tumor core segmentation result will be used to segment enhancing tumor. We not only use the final output of the model, but also combine the results of intermediate output. In MICCAI BraTs 2018 gliomas segmentation task, we achieve a competitive performance without data augmentation.

Keywords: 3D-SE-Inception-ResNet · Cascaded FCN · Anisotropic · Medical image segmentation

1 Introduction

Image segmentation has always been a challenging task in the field of computer vision. Especially in medical image field, multi-modal Magnetic Resonance Images can be used to segment human body pathological tissue. Many medical committees such as MICCAI, have always been focusing on the evaluation of state-of-the-art methods for the segmentation of brain tumors in multi-modal Magnetic Resonance Imaging (MRI) scans. In 2D image processing fields, many effective models were proposed. AlexNet, presented by Krizhevsky et al. [12], won the image classification task of ImageNet 2012. Since then the method of deep learning has aroused researchers' attention. Later, Deep learning models have been kept explosive growth. VGGNet [17], used a series of small convolution filters to substitute for large convolution filters. GoogleNet [19] proposed

A. Crimi et al. (Eds.): BrainLes 2018, LNCS 11384, pp. 346–357, 2019.
https://doi.org/10.1007/978-3-030-11726-9_31

a multi scale concept, by using different size filters to extract information, and its improved version Inception [18], creatively used $1 * 1$ convolution filters to reduce the number of model parameters, while ensuring the model depth without increasing the parameters of the model. Squeeze and Excitation Networks [11], a kind of attention mechanism, introduced the attention mechanism into the spatial dimension, further improving the performance of the model. However, using multi-modal Magnetic Resonance Images to segment human tissue has been very challenging. Because medical image data is more complex than ordinary image data, both plane information and spatial information should be considered. So some researchers try to solve the problem of medical image segmentation by using deep learning method. In the first attemp, the modified variants of 2D CNN was adopted, by using aggregated adjacent slices [6] or orthogonal planes [15,16], but this method did not take into account space information, it couldn't segment object accurately. Recently, a variety of 3D models had been developed to segment objects from volumetric data and gained competitive performance. For examples, 3D U-Net [8] allows end-to-end training and testing for volumetric image segmentation. VoxResNet [5], a deep voxelwise residual network, improves the volumetric segmentation performance by seamlessly integrating the low-level image appearance features, implicit shape information and high-level context together.

The contribution of this paper are four-fold. First, we combine the mainstream segmentation models of 2D CNNs [13] and modified Inception structure to deal with 3D images. In the process of designing the model, we also consider the computation performance, and design two kinds of Inception layer, which are named as Lower Inception and Higher Inception. Second, we apply the 3D Squeeze and Excitation structure to our model. Third, we use multi-scale filters to downsample the 3D feature maps, the loss of valid information can be better reduced when resizing the 3D feature maps. Fourth, our model uses the residual connection to make sure the information can be transferred better and the training process of the model can be accelerated.

2 Methods

2.1 Cascaded Framework

The cascaded framework [7,22] is designed to simplify segmentation problems. We use triple cascaded networks to segment substructures of brain tumor, each network can be seen as a binary segmentation network. While the first network segments the whole tumor task according to the MRI, a bounding box of the whole tumor is obtained. The region of the input images is cropped based on the bounding box, and the cropped result is used as the input of the second network to segment tumor core. After segmenting tumor core, another smaller bounding box is obtained. The image region is resized according to the smaller bounding box of the tumor core. Then the resized image region is used as the input of the third network to segment the enhancing tumor core. During the training phase, the bounding boxes are decided by the ground truth. In the testing stage, the bounding boxes are generated based on the segmentation results.

2.2 Neural Networks Architecture

The overall architecture of the model we proposed is shown in Fig. 1. It includes inception layers, SE structures, reduction layers and residual connection. High-Level Inception uses dilated convolution. The model contains a great deal of 2D image mainstream model structures. Considering the huge advantages of their own structures in 2D images, modifying them to adapt the 3D medical image data can have better effects.

Low-Level Inception. The Low-level Inception structure is shown in Fig. 2. We use $1 * 3 * 3$, $1 * 5 * 5$ and $1 * 7 * 7$ convolution filters to better capture the information of feature maps early in the networks. Why do we design model like this? There are several model design principles [20]. The first principle is to avoid representational bottlenecks, especially early in the network. Any feed-forward networks can be seen as an acyclic graph from input to output. Once the model is defined, the flow direction of information will be decided. When the information passes the model, information is fading. We use the multi large receptive fields early in the network to avoid bottlenecks with extreme compression. In AlexNet [12], Krizhevsky et al. used the $11 * 11$ receptive fields. However, large convolution filters have a serious shortcoming, i.e., large convolution filters have a huge number of training parameters. The parameters of $7 * 7$ receptive fields are 5 times as much as those of $3 * 3$ receptive fields. But large receptive fields can better capture the space information. We should consider the trade-off between computation performance and model complexity, so we apply the large convolution filters only to the first four layers. We use different multi-scale size filters to better capture the space information, while avoiding representation bottleneck.

3D Squeeze and Excitation Structure. Squeeze and Excitation structure was proposed by Hu et al. [11] in 2017, they used the SENet to get a top performance in the ImageNet 2017. The innovation of this model is to explicitly model the interdependence between feature channels. Specifically, it is important to acquire each characteristic channel automatically through the way of learning, improve the useful features and restrain the small features of the current task in accordance with its importance. Based on this idea, we redefine the squeeze and excitation operation in our model. For any given transformation $F_{tr} : X \rightarrow U, X \in \mathbb{R}^{D' \times W' \times H'}, U \in \mathbb{R}^{D' \times W' \times H'}$. We take F_{tr} as a standard 3D convolution operator. $V = [V_1, V_2, ..., V_C]$ denotes the learned set of filter kernels, where V_C refers to the parameters of the c-th filter. We denote $U = [u_1, u_2, ..., u_c]$ as the output of F_{tr}, where

$$u_c = v_c * X = \sum_{s=1}^{C'} v_c^s * x^s \tag{1}$$

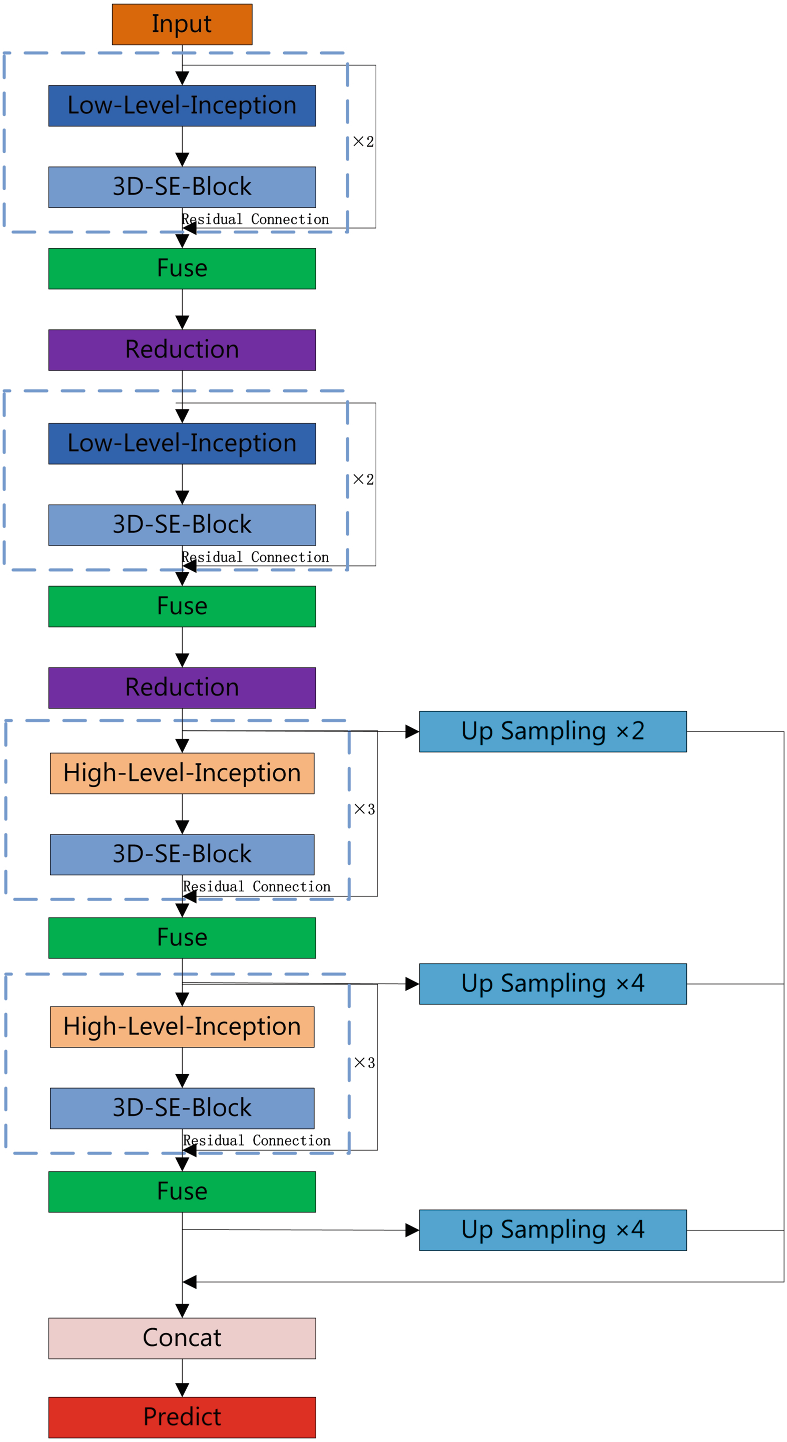

Fig. 1. The architecture of the model we proposed.

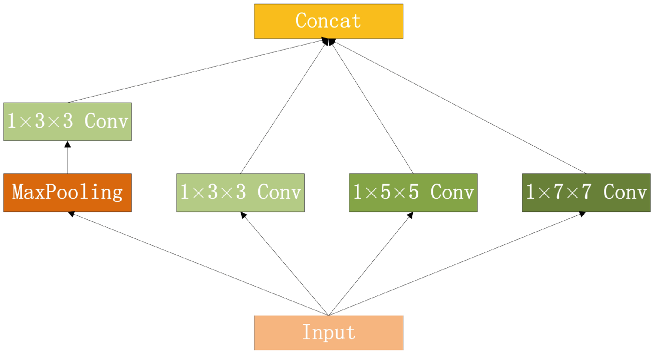

Fig. 2. The architecture of Low-Level Inception

Here $*$ denotes convolution operation, $v_c = [v_c^1, v_c^2, ...v_c^{c'}]$ and $X = [x^1, x^2, ...x^{c'}]$, v_c^s is a 3D spatial kernel, and represents a single channel of v_c, which acts on the corresponding channel of X.

3D Squeeze: We perform feature compression along the space dimension, turning each of the three dimensional characteristic channels into a real number. This real number has a global receptive field to some extent, and the output dimension matches the number of input characteristic channels. It represents the global distribution of responses on characteristic channels. Moreover, the whole receptive field can be obtained near the input layer.

$$z_c = F_{sq}(u_c) = \frac{1}{D \times W \times H} \sum_{i=1}^{D} \sum_{j=1}^{W} \sum_{k=1}^{H} u_c(i, j, k) \tag{2}$$

Here, a statistic $z \in \mathbb{R}^c$ is generated by shrinking U through spatial dimensions $D \times W \times H$, z_c denotes the c-th element of z.

3D Excitation: Excitation operation is a mechanism similar to recurrent neural network's middle gate. Parameters are used to generate weights for each characteristic channel, the parameters are learned to explicitly model the correlation between feature channels. Then, we use the sigmoid activation as a simple gating mechanism:

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2\delta(W_1 z)) \tag{3}$$

where δ refers to the *ReLu* function [13], $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ and $W_2 \in \mathbb{R}^{C \times \frac{C}{r}}$. After *ReLu* function, we add two fully-connected layers to limit model complexity. r denotes the reduction ration.

Output: The final output is a reweight operation. It's obtained by rescaling the transformation output U with the activations:

$$\widetilde{x_c} = F_{scale}(u_c, s_c) = s_c \cdot u_c \tag{4}$$

where $\widetilde{X} = [\widetilde{x_1}, \widetilde{x_2}, ..., \widetilde{x_c}]$, and $F_{scale}(u_c, s_c)$ refers to channel-wise multiplication between the feature map $u_c \in \mathbb{R}^{D\times W\times H}$ and the scalar s_c. 3D SE structure is a kind of attention mechanism that can pay attention to 3D channels relationship. c denotes the channels, r denotes ration (In our model, the ration $r = 4$, 8 and 16, we tested the model separately). The SE structure is shown in Fig. 3.

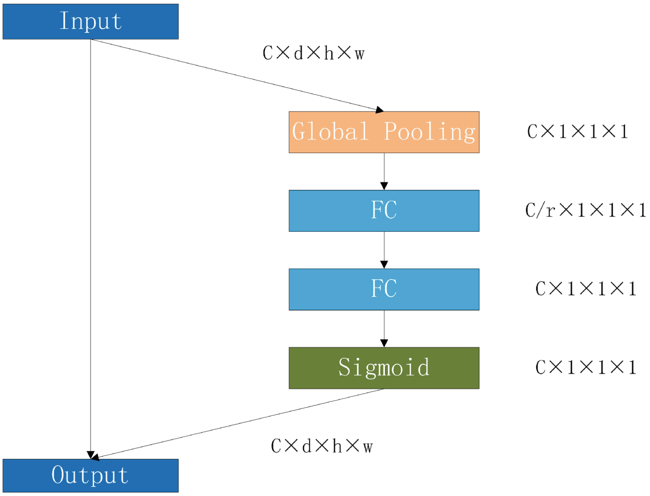

Fig. 3. The architecture of 3D squeeze and excitation

Reduction Structure. Reduction structure is used for reduction feature maps. As mentioned before, using multi-scale can capture more spatial information. Different variants of this blocks (with various number of filters) can be set by users, here we set the number of m, n, o, k and I as 8. As shown in Fig. 4, reduction structure can use multi-scale convolution to capture the information from input feature maps. m, n, o, k and i can be set arbitrarily. We consider the simplified model, so set all the variables to the same number 8 and use $1*1*1$ convolution. In the design principles we mentioned earlier [20], the second principle is intent to let the spatial aggregation be done over lower dimensional embeddings without affecting representational power. Considering that these signals are easy to be compressed, dimensionality reduction will speed up the learning process. We redesign the reduction structure according to this idea.

High-Level Inception. The High-level Inception structure is shown in Fig. 5. The third principle is to factorize a large convolution kernel into smaller ones. Convolutions with large filters have a huge computation complexity. For example, in the case of the same number of convolution kernel, $1 * 5 * 5$ convolution is $25/9 = 2.78$ times more computationally complex than that of $1 * 3 * 3$. But simply reducing the size of the convolution core will cause information loss.

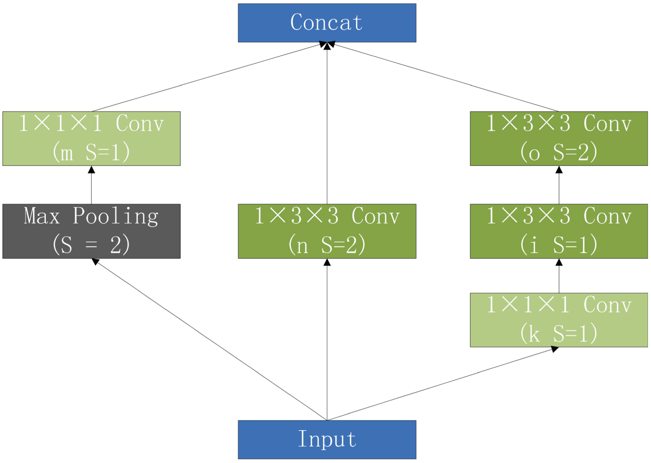

Fig. 4. The architecture of reduction module

However 1 ∗ 5 ∗ 5 convolution can be replaced by multi-layer small convolution networks. Look at the 1 ∗ 5 ∗ 5 network as full convolution, each output is a convolution kernel slipping on the input, it can be replaced by two 1 ∗ 3 ∗ 3 convolutional layer. The convolution of High-Level Inception uses the dilated convolution kernels. The dilated convolution uses small filters but has a larger receptive fields, without increasing the parameters. We set the dilation rate 1, 2, 3 and 3, 2, 1 corresponding to each High-Level Inception layers in order. The High-Level Inception architecture we designed can be seen in Fig. 5.

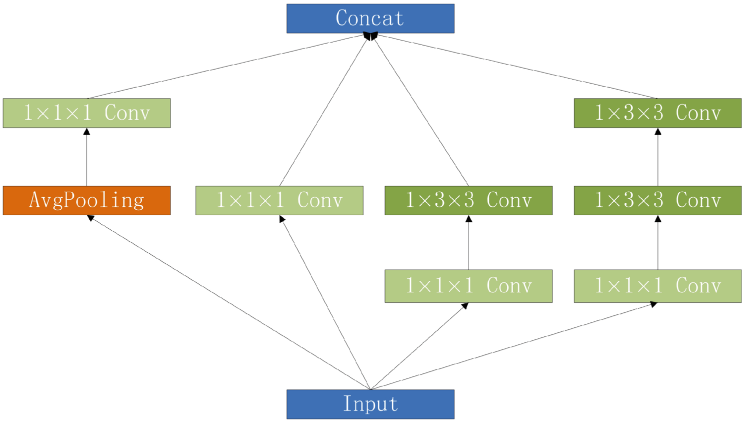

Fig. 5. The architecture of High-Level Inception

Residual Connection. ResNet was put forward in 2015 by He et al. [10], it won the first place in the classification competition of ImageNet. With the increasing of network depth, the problem of the disappearance of the gradient is becoming more and more obvious. The training of the network has become quite difficult. The basic idea of ResNet is to introduce "shortcut connection" that can skip one or more layers. ResBlock can be defined as:

$$y = F(x, w_i) + x \tag{5}$$

Here x and y are the input and output vectors of the layers considered. The function $F(x, w_i)$ represents the residual mapping to be learned. If the dimensions of x and F don't equal, we can perform a linear projection W_s by the shortcut connections to match the dimensions:

$$y = F(x, w_i) + W_s x \tag{6}$$

W_s is used only when matching dimensions.

Prediction and Fusion. In the prediction phase, we not only use the final result but also use the intermediate output results, and concatenate them as the final prediction result. In the training phase, each neural network is trained in axial, sagittal and coronal views. During the test phase, predictions are fused to get the final segmentation. We average the softmax outputs in these cascade networks. Fusion structure is a simple $3 * 1 * 1$ convolution, as one can see the green block in the Fig. 1. The overall model decomposes $3 * 3 * 3$ convolution kernels to $1 * 3 * 3$ convolution and $3 * 1 * 1$ convolution, $1 * 3 * 3$ convolutions are used to extract the datasets features and $3 * 1 * 1$ convolutions are used to fusion the datasets spatial features.

3 Experiments and Results

Brain tumor segmentation is a challenging task, which has attracted a lot of attentions in the past few years. We use the BRATS 2018 dataset [1,2], which is composed of multiple segmentation subproblems. The whole tumor region is identified in a set of multi-modal images, tumor core areas and active tumor regions [4,14].

Medical Image Data. Brats 2018 dataset contains real volumes of 210 high-grade and 75 low-grade glioma subjects. For each patient, T1Gd, T1, T2, FLAIR and Ground Truth MR volumes are available. These 285 subjects are used in training set, and there are 66 other subjects as the validation dataset. Considering the unbalance distribution of the training data, we expand the LGG dataset 3 times based on the original one, during the training data loading process, each LGG dataset copies and reloads 3 times. When training the network, we randomly choose 5 subjects as the input. All of these volume average size is

$155 * 240 * 240$, we resize the volume and extract the voxel of specified shape in the middle volume as the final training input. The biggest black box outside represents the source MRI data set, and the middle gray bounding box represents the valid volumes (include human brain tissue), red point is the core of the target size train patch. In the valid volumes bounding box, dotted line box random crops with the center of the red point, it's used to train our neural networks, we train three cascaded anisotropic networks, use the different patch size to train the different network. We extract the (26, 120, 120) patch size for training the whole tumor segmentation network, (26, 72, 72) patch size for training the tumor core segmentation network and (26, 48, 48) patch size for training the enhancing tumor segmentation network. The details are shown in Fig. 6.

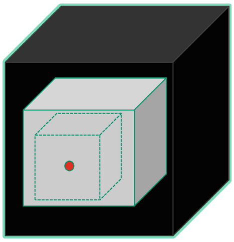

Fig. 6. Data preprocessing details sketch map (Color figure online)

Training Details. Our network is implemented in Tensorflow and NiftyNet, no external data was used during the training. We use Adam optimizer to train, and use PReLu [9] as the activation, set the batch-size = 5, weight decay = 10^{-7}, learning rate = 10^{-3}, max-iteration = $20k$. We train on the GTX 1080Ti GPU. For the data pre-processing, the images are normalized by the mean and standard deviation. And we use the Dice coefficient as the model loss function.

Segmemtaion Results. In order to test the influence of parameter r on the model, we perform three groups of experiments. However, the experiments show that too large or too small parameter r can not get the best result, a moderate parameter r can achieve a better result. More details are shown in Tables 1, 2 and 3. The result of Table 2 is the best among all of them. From the perspective of SE structure, parameter r relates to the number of the first fully connected layer ($fc = c/r \times 1 \times 1 \times 1$), when we give the parameter r a small number, the number of the first FC layer will be quite large, it will increase computation complexity, makes the model hard to train, as shown in Table 1. But if we set the parameter a large number, the number of the first FC layer will be small, it will make the model difficult to learn the channel characteristics better, as can be seen in Table 3. At present, there is no authoritative idea on how to set the parameter r. You can only adjust the parameter r according to the result of experiments. It is regrettable that we failed to submit our best results before the deadline. Table 4 shows the our official scores computed by the organizer of the challenge. Besides, we also test our model on the Brats 2015 dataset with good results. The detail results of our model are shown as Table 5.

Table 1. Table shows the result of our model predict (ration = 4).

Data Set	Dice			Sensitivity			Specificity			Hausdorff95		
	WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET
Training	0.729	0.885	0.834	0.805	0.924	0.879	0.998	0.992	0.996	5.657	16.999	7.082
Validation	0.784	0.878	0.807	0.814	0.935	0.844	0.998	0.991	0.997	4.380	19.034	9.408

Table 2. Table shows the result of our model predict (ration = 8).

Data Set	Dice			Sensitivity			Specificity			Hausdorff95		
	WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET
Training	0.773	0.910	0.872	0.832	0.929	0.886	0.998	0.994	0.997	3.738	6.938	4.644
Validation	0.798	0.901	0.813	0.818	0.933	0.831	0.998	0.993	0.997	4.158	6.371	8.840

Table 3. Table shows the result of our model predict (ration = 16).

Data Set	Dice			Sensitivity			Specificity			Hausdorff95		
	WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET
Training	0.768	0.910	0.868	0.822	0.918	0.879	0.998	0.994	0.997	3.974	6.878	4.841
Validation	0.796	0.903	0.818	0.810	0.928	0.820	0.998	0.993	0.998	3.971	6.255	8.371

Table 4. Performance of proposed method on Test Dataset (model ration = 4).

Label	Dice-ET	Dice-WT	Dice-TC	Hausdorff95-ET	Hausdorff95-WT	Hausdorff95-TC
Mean	0.724	0.864	0.772	5.353	9.131	8.115
StdDev	0.277	0.138	0.263	10.431	14.717	12.041
Median	0.828	0.909	0.889	2.236	3.317	3.742
25quantile	0.710	0.857	0.727	1.414	2.236	2
75quantile	0.879	0.938	0.928	3.317	6.164	8.108

Table 5. Brats 2015 test set results: we rank 8th on the Brats 2015 leaderboard.

Data Set	Dice			Positive			Sensitivity			Rank		
	WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET
Test	0.86	0.73	0.63	0.85	0.82	0.61	0.89	0.71	0.70	14.25	15.25	32.50

4 Conclusion

The results of all participants can be seen in [3], compared with other participants, our results can achieve more competitive performance than many of them. As is shown above, setting the model parameter r = 8 can achieve better results than others. We don't perform enough parameter adjustment experiments and don't use other optimization algorithms. When processing data, we only use single volume size. In the future, we plan to integrate convolution CRFs [21] and self-attention of 2D image segmentation into our model.

Acknowledgements. We'd like to thank the team of Wang [22] for their source code, with which we can focus more on the innovation of the model and do our work more easily. And we also thank the NiftyNet team, the deep learning tools they developed enables us construct our model more efficiently. This work was supported by the Natural Science Foundations of China under Grants No. 61463050, No. 61702443, No. 61762091, the NSF of Yunnan Province under Grant No. 2015FB113, the Project of Innovative Research Team of Yunnan Province under Grant No. 2018HC019.

References

1. Bakas, S., et al.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection. Cancer Imaging Arch. (2017). https://doi.org/10.7937/K9/TCIA.2017.KLXWJJ1Q
2. Bakas, S., et al.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection. Cancer Imaging Arch. (2017). https://doi.org/10.7937/K9/TCIA.2017.GJQ7R0EF
3. Bakas, S., Reyes, M., et al.: Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv preprint arXiv:1811.02629 (2018)
4. Bakas, S., et al.: Advancing the Cancer Genome Atlas Glioma MRI collections with expert segmentation labels and radiomic features. Sci. Data **4**, 170117 (2017)
5. Chen, H., Dou, Q., Yu, L., Heng, P.A.: Voxresnet: deep voxelwise residual networks for volumetric brain segmentation. arXiv preprint arXiv:1608.05895 (2016)
6. Chen, H., Yu, L., Dou, Q., Shi, L., Mok, V.C., Heng, P.A.: Automatic detection of cerebral microbleeds via deep learning based 3D feature representation. In: 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), pp. 764–767. IEEE (2015)
7. Christ, P.F., et al.: Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3D conditional random fields. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 415–423. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_48
8. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49

9. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1026–1034 (2015)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
11. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. arXiv preprint arXiv:1709.01507 7 (2017)
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
13. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
14. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging **34**(10), 1993 (2015)
15. Prasoon, A., Petersen, K., Igel, C., Lauze, F., Dam, E., Nielsen, M.: Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013. LNCS, vol. 8150, pp. 246–253. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40763-5_31
16. Roth, H.R., et al.: A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014. LNCS, vol. 8673, pp. 520–527. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10404-1_65
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
18. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI, vol. 4, p. 12 (2017)
19. Szegedy, C., et al.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
20. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
21. Teichmann, M.T., Cipolla, R.: Convolutional CRFs for semantic segmentation. arXiv preprint arXiv:1805.04777 (2018)
22. Wang, G., Li, W., Ourselin, S., Vercauteren, T.: Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks. In: Crimi, A., Bakas, S., Kuijf, H., Menze, B., Reyes, M. (eds.) BrainLes 2017. LNCS, vol. 10670, pp. 178–190. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75238-9_16