AIU-Net: An Efficient Deep Convolutional Neural Network for Brain Tumor Segmentation

Yongchao Jiang,1,2 Mingquan Ye,1,2 Daobin Huang,1,2 and Xiaojie Lu1,2

1School of Medical Information, Wannan Medical College, Wuhu 241002, China
2Research Center of Health Big Data Mining and Applications, Wannan Medical College, Wuhu 241002, China

Correspondence should be addressed to Mingquan Ye; ymq@wnmc.edu.cn

Received 20 May 2021; Revised 23 July 2021; Accepted 29 July 2021; Published 4 August 2021

1. Introduction

Glioma is the most common brain tumor, and it is also the brain tumor with the highest mortality and morbidity. Accurate segmentation of gliomas is of great significance for the diagnosis and treatment of gliomas. Magnetic resonance imaging (MRI) is an important technical means to assist doctors in the diagnosis and treatment of brain tumors [1]. Various sequences of MRI can provide different brain tumor tissue structures, and it is usually combined with multimodal MRI of brain tumor to segment brain tumors. Because of the complexity of brain tumor structure, the fuzziness of tumor boundary, and the difference of different individuals, the accurate segmentation of brain tumor is a complicated and difficult task [2]. The traditional manual segmentation needs a lot of time for doctors to complete, and the segmentation accuracy is relatively rough. In recent years, the automatic segmentation method based on deep learning has achieved good results in medical image segmentation [3].

The deep learning method based on convolutional neural network performs well in computer vision tasks such as image classification, segmentation, and object detection. Convolutional neural network can automatically learn the complex features of data in the training process without relying on manual extraction of features, which can further improve the segmentation accuracy of brain tumors [4, 5]. Long et al. [6] proposed the full convolutional neural network (FCN), which transformed the full connection layer of the convolutional neural network into the convolutional layer and used up sampling to restore the output feature map to the same size as the input image, so as to achieve end-to-end semantic segmentation of the image. Ronneberger et al. [7] proposed U-net for biological cell segmentation; U-net is composed of contracting path and expanding path.
Contracting path includes convolution block for feature extraction and max-pooling for down sampling. Expanding path is composed of convolution block and up-sampling module. Features with the same resolution are fused by skip connection between contracting path and expanding path. U-net has a simple structure and can obtain better segmentation results in the case of small sample size of medical images. However, the application of U-net in brain tumor image segmentation still needs to be improved [8]. On the one hand, the contracting path of U-net uses the pooling layer to reduce the feature map to expand the receptive field. Continuous pooling operation may cause the loss of image details and affect the segmentation result. On the other hand, it is the difference of the size, shape, and location of brain tumor, and how to obtain more detailed features of segmentation target and how to obtain multiscale features are the important problems [9]. In order to solve the problem of gradient disappearance and network degradation with the increase of network depth, the residual network (ResNet) [10] is proposed. By adding identity mapping between the input and output of several convolutional layers, the network is easier to converge and prevent network degradation [11].

In order to obtain multiscale features, Chen et al. [12] proposed the DeepLab model. In this model, the last pooling layers were removed, and atrous convolutions were used to expand the receptive field. The atrous spatial pyramid pooling (ASPP) samples the given input in parallel with the atrous convolution of different dilation rates and then splices the results together. ASPP has a better effect on the extraction of multiscale features. Szegedy et al. [13] proposed inception network. Inception module can increase the width of the network. GoogLeNet, which is composed of inception module, obtained the best classification and detection performance in ILSVRC 2014 competition. This paper proposes to use atrous convolution to expand the receptive field and reduce the use of pooling layer, so as to reduce the loss of image details. Atrous convolution is to insert holes into the standard convolution kernel to expand the receptive field of feature extraction without additional parameters [14]. In this paper, atrous convolution and inception are combined to form a new structure named A-Inception module, and a new network architecture based on U-net is proposed. The encoder of this network adopts A-inception module to increase the depth and width of the network and obtain the different sizes of receptive field. At the same time, the atrous spatial pyramid pooling is added into the network to extract the multiscale features of the image.

2. Related Work

In recent years, methods based on convolutional neural networks have provided good performance in the field of computer vision. Compared with traditional methods, algorithms based on convolutional neural network can automatically learn the complex features of the original data without relying on manual extraction of features, which further improves the accuracy of image segmentation [15].

The framework of encoder-decoder is a common structure in image segmentation. In the encoding process, the pixels of the image are mapped to a high-dimensional distribution, and the decoding process is to gradually restore the details and spatial dimensions of the image. Therefore, the encoder-decoder structure can achieve the end-to-end semantic segmentation of the image [16]. SegNet [17] is a typical encoder-decoder network framework in image segmentation. The encoder network in SegNet has the same topology as the convolution layer in VGG16, but removes the fully connected layers. The network uses max-pooling to reduce the dimension of feature maps, and the decoder uses max-pooling indices received from the corresponding encoder to perform nonlinear up sampling of their input feature maps. U-net is also encoder-decoder structure and has been widely used in medical image segmentation. It adds skip connections between the encoder and the decoder, which are used to fuse the feature maps with the same resolution between the encoder and the decoder. Shaikh et al. [18] introduced dense connection and replaced the basic convolution module in U-net with dense connection module, which further improved the segmentation performance of the network. Oktay et al. [19] introduced the attention gates into the standard U-net architecture that automatically learns to focus on target structures of varying shapes and sizes. In the training process, the attention weight gradually tended to the target region, while the attention weight of the background region gradually decreased so that the segmentation accuracy was improved.

In the semantic segmentation of images, convolutional neural network uses pooling to realize down sampling, which reduces the image size and increases the receptive field and then uses up sampling to restore the original image size. In this process, some detail features of the image will be lost [20]. Atrous convolution can increase the receptive field without losing the image resolution, thus improving the accuracy of image semantic segmentation. Zhao et al. [21] proposed pyramid scene parsing network (PSPNET), which aggregates the context of different regions through pyramid pooling module and improves the ability of the network to obtain global information. In order to segment multiscale objects, DeepLabv3 [22] proposes to connect several atrous convolutions with different dilation rates in series and parallel, which can obtain larger receptive fields in cascade mode, and different receptive fields in the parallel mode for the same input, which can extract multiscale features better. DenseASPP [23] integrates atrous convolutions with different dilation rates through dense connection. Without the use of pooling operation, the receptive field of output neurons is expanded so that the output features cover a large range of semantic information and acquire multiscale features. DeepLabv3+ [24] is an extension of DeepLabv3, adding a simple decoder module to recover the object boundary details. The ASPP is an improvement on the basis of spatial pyramid pooling. For multiscale object segmentation, parallel pooling modules are designed to obtain multiscale features.
3. Methods

3.1. Atrous Convolution. When convolutional neural network is used for end-to-end semantic segmentation of images, down sampling will reduce the resolution of the feature maps, which can reduce the amount of computation and expand the receptive field. After that, the feature maps can be restored to the original image size through up sampling. In this process, some details related to the boundary of the segmentation object will be lost, resulting in the image segmentation results which are not accurate enough. Atrous convolution can control the resolution of features and adjust the size of receptive field to capture multiscale information. In fact, atrous convolution is to inject holes into the standard convolution kernel, and the dilation rate is used to define the interval of convolution kernel insertion [25]. The atrous convolution with the dilation rate of 1 is the same as the standard convolution. Figure 1 shows the atrous convolution with dilation rates of 1, 2, and 3, respectively. Compared with the ordinary convolution with convolution kernel of $3 \times 3$, the receptive field of atrous convolution is larger.

Increasing the depth and width of the network to improve network performance will bring a large number of parameters, which can easily lead to overfitting and increase the amount of calculation. The fundamental method to solve this problem is to keep the sparsity of neural network structure, but the computational efficiency of computer for nonuniform sparse data is very low. A large number of literatures show that the sparse matrix can be clustered into relatively dense submatrix to improve the computational performance. The main purpose of the inception structure is to use dense components to approximate the optimal local sparse structure. In this paper, a new module A-Inception is proposed. In this module, there are three branches in parallel, each branch uses different convolution kernel, instead of directly connecting convolution kernel in series, thus increasing the width of the network. In this module, atrous convolution is used to replace ordinary convolution. Different branches have different receptive fields. The convolution of different receptive fields is connected in parallel. Because the size of brain tumors is greatly different, receptive fields of different scales can reduce the fluctuation caused by the disturbance of brain tumor size, improve the robustness of neural networks, and obtain more detailed features at the same time. The BN layer is added after each convolution layer to avoid the gradient vanishing [26]. At the same time, inspired by the Inception-ResNet [27] module, the residual connection is added between the input and the output, which makes the network easier to learn and faster to converge. The specific model structure is shown in Figure 2.

3.2. ASPP. In Deeplabv2, atrous spatial pyramid pooling is proposed, which uses atrous convolutions with different dilation rates in parallel to obtain multiscale features of images. In Deeplabv3, the BN layer is added to the atrous spatial pyramid pooling, and global pooling is paralleled. Atrous convolutions with different dilation rates have different receptive fields for the same input, and these results can be stitched together to better capture the multiscale features of the image. In order to reduce the number of channels after splicing, the $1 \times 1$ convolution layer is connected. In [22], when the output-stride = 16 (output-stride is the ratio of input image spatial resolution to final output resolution), three $3 \times 3$ convolutions with rates = 6, 12, 18 are adopted in ASPP, while when the output stride = 8, the rates should be doubled. In this paper, three times downsampling is used in the encoder part, so the output stride is 8. The experiment proves that the segmentation effect is better when the atrous convolutions of $3 \times 3$ with rates = 12, 18, 24 are used. The ASPP module adopted in this paper is shown in Figure 3. In addition to three atrous convolutions with different rates, there is also a $1 \times 1$ convolution and a pooling layer in parallel.

3.3. Network Structure. In this paper, an improved brain tumor segmentation algorithm based on U-net is proposed. The encoder obtains the higher-level semantic information of the image, and the decoder gradually recovers the spatial information of the image. The encoder uses five A-inception modules. The first three A-inception modules use the atrous convolution with the dilation rate of 1, which is the standard $3 \times 3$ convolution, and then use the down-sampling module to reduce the feature resolution. The down-sampling modules use the max-pooling and the $3 \times 3$ convolution with the stride of 2 for down sampling the input, respectively, and then parallel the results of the two. In order to reduce the use of the pooling layer and prevent more loss of image details, the last two A-inception modules in the encoder use atrous convolution with larger dilation rates. The rate1, rate2, and rate3 are 2, 2, 4 and 4, 4, 8 in A-inception block 4 and A-inception block 5, respectively. It cannot only expand the receptive field but also connect different receptive fields in parallel, which can better capture multiscale features and obtain more image details [28]. The ASPP module is used between the encoder and decoder. The decoder uses three residual blocks and bilinear interpolation up sampling to restore the feature maps to the same size as the input image. At the same time, the feature maps of the same resolution in the encoder and decoder are combined, and low-level features are introduced to increase the segmentation accuracy of spatial information features. The specific network structure is shown in Figure 4:

The optimization algorithm adopted in this paper is adaptive moment estimation (Adam) [29], which has the advantages of simple implementation, low memory requirement, and high computational efficiency. The loss function used in this paper is a linear combination of cross-entropy loss function and Dice loss function. The Dice loss function is suitable for the situation where the positive and negative samples are not balanced, and it focuses more on the prospects. However, if there are many small targets in the training data during the experiment, the loss curve is likely to oscillate. Therefore, this paper adopts the loss function of the Dice loss function combined with the cross-entropy loss.
function, which can alleviate the problem of sample imbalance and obtain a smoother loss curve. The loss function is defined as follows:

\[
\text{Loss}_{\text{CE}} = -\sum_{i} g_i \log(p_i),
\]

\[
\text{Loss}_{\text{Dice}} = 1 - \frac{2 \times \sum_{i} p_i g_i}{\sum_{i} p_i + \sum_{i} g_i},
\]

\[
\text{Loss} = 0.5 \text{Loss}_{\text{CE}} + \text{Loss}_{\text{Dice}}.
\]

The set \(N\) of all samples is calculated, where \(g_i\) is the thermal code (0 or 1) of the \(i\)th sample tag and \(p_i\) is the prediction probability of the \(i\)th sample tag.

3.4. Data Processing. The experimental data used in this paper are brain tumor challenge datasets BraTS2018 and BraTS2019. BraTS2018 dataset includes 210 HGG patients and 75 LGG patients, each of which includes T1 (T1 weighted), T2 (T2 weighted), T1c (contrast enhanced T1 weighted), and flair (fluid attenuated inversion recovery) four MRI sequences and ground truth labels [30]. These data were used as a training set for the experiment. BraTS2019 dataset added 49 HGG patients and 1 LGG patient on the basis of BraTS2018 dataset, and these data were used as the testing set of the experiment. The size of each modal MR image is 240 \(\times\) 240 \(\times\) 155. The ground truth labels are the result of tumor manually labeled by 1 to 4 experts according to the same annotation protocol, including normal tissue (label 0), necrotic and nonenhancing tumor (label 1), edema (label 2), and enhancing tumor (label 4) [31].

In the data preprocessing method, firstly, the data are standardized, that is, subtracting the mean value and dividing by the standard deviation. Then, the redundant background in the original data is cropped to alleviate the problem of data imbalance [32]. Then, the three-dimensional images were sliced to obtain two-dimensional images, and the slices without lesions in the training set were discarded to alleviate the category imbalance. The slices of four MRI scan modalities are combined into four channel samples for training and testing data.

4. Results and Discussion

The experimental environment is Intel Xeon Silver 4116 CPU@2.10 GHz, GPU NVIDIA GeForce RTX2080Ti. After preprocessing the experimental data, a two-dimensional image sample with the size of 160 \(\times\) 160 is obtained. 80% of the training set is used for model training, and 20% is used as the validation set to adjust parameters to monitor whether the model is overfitting. The testing set is used to verify the segmentation effect.

4.1. Evaluation Metrics. In order to quantitatively evaluate the segmentation performance of the proposed algorithm, the evaluation metrics used in this paper include the Dice similarity coefficient (DSC), intersection over union (IOU), and positive predictive value (PPV). These indicators were used to evaluate the experimental results [33]. The Dice similarity coefficient represents the similarity between the experimental segmentation results and the ground truth labels. DSC, PPV, and IOU are commonly used in image segmentation. The definitions are as follows:

\[
\text{DSC} = \frac{2 \times \text{TP}}{\text{FP} + 2 \times \text{TP} + \text{FN}},
\]

\[
\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}},
\]

\[
\text{IOU} = \frac{\text{TP}}{\text{FP} + \text{TP} + \text{FN}},
\]

where TP is true positive, FP is false positive, and FN is false negative. The range of the result is 0 to 1. The closer the test result is to 1, the more accurate the segmentation result is.

4.2. Experimental Results. In this paper, the results of glioma segmentation include the whole tumor region and tumor subregion, which are whole tumor region (WT), tumor core (TC), and enhancing tumor region (ET). ET is enhancing tumor, TC includes enhancing tumor and necrotic, and WT includes regions of enhancing tumor, necrotic, and edema. In this paper, an improved network architecture AIU-net
based on U-net is proposed. In order to test the segmentation effect of this network on brain tumors and compare it with U-net and DeepLabv3+ network structure, the same dataset and parameters are used for training and testing on each network. The experimental results are shown in Figure 5. From the segmentation results shown in the figure, it can be seen that the method proposed in this paper is better than the other two methods in the segmentation of tumor details, and the segmentation results are closer to the ground truth. Figure 6(a) shows the change of the loss of the three networks with epoch in the training process. It can be seen from the figure that when epoch was 50, the IOU value of AIU-net had exceeded 0.9, while the IOU value of the other two networks were all less than 0.9.

Table 1 shows the evaluation results of DSC, IOU, and PPV of several methods. The results show that the values of DSC, IOU, and PPV of the proposed method are higher than those of U-net and DeepLabv3+, and the values of DSC, IOU, and PPV of the whole tumor region are greatly improved compared with those of the other two methods. The segmentation results of the whole tumor region, enhancing tumor region, and tumor core region are better than those of the other two methods, and the segmentation performance is improved.
Figure 5: Example segmentation results on the BRATS dataset. From left to right are the segmentation results of U-net, DeepLabv3+, AIU-net, and ground truth. The whole tumor (WT) class includes all visible labels (a union of green, yellow, and red labels), the tumor core (TC) class is a union of red and yellow, and the enhancing tumor core (ET) class is shown in yellow. (a) U-net. (b) DeepLabv3+. (c) AIU-net. (d) GT.

Figure 4: Overall architecture of the AIU-net model.
5. Conclusion

In order to improve the accuracy of brain tumor automatic segmentation, this paper proposes a network architecture AIU-net based on U-net, which uses a new module combining inception and atrous convolution as encoder, and introduces ASPP module to obtain multiscale features. Experiments show that the new architecture AIU-net can effectively improve the accuracy of brain tumor segmentation and is conducive to multiscale information extraction. The segmentation of tumor details has been improved. In comparison with U-net and DeepLabv3+, the results of DSC, IOU, and PPV of the proposed method are better than those of other methods, and better segmentation performance is obtained. However, compared with U-net and DeepLabv3+, the training time and test time of the proposed method are longer mainly because the network architecture is more complex. The further work in the future is to obtain higher segmentation accuracy and better efficiency at the same time.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was funded and supported in part by the National Natural Science Foundation of China, under Grant 61672386, and the Anhui Provincial Natural Science Foundation of China, under Grant 1708085MF142.

References

[1] D. Ivana, G. Bart, and P. Wilfried, "MRI segmentation of the human brain: challenges, methods, and applications," *Computational and Mathematical Methods in Medicine*, vol. 2015, Article ID 450341, 23 pages, 2015.

[2] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: achievements and challenges," *Journal of Digital Imaging*, vol. 32, no. 8, pp. 582–596, 2019.

[3] A. İşın, C. Direkoğlu, and M. Şah, "Review of MRI-based brain tumor image segmentation using deep learning methods," *Procedia Computer Science*, vol. 102, pp. 317–324, 2016.

[4] P. Moeskops, M. A. Viergever, A. M. Mendrik et al., "Automatic segmentation of MR brain Images with a
convolutional neural network,” *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1252–1261, 2019.

[5] Y. Zhang, C. Yin, Q. Wu, Q. He, and H. Zhu, “Location-aware deep collaborative filtering for service recommendation,” *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 6, pp. 3796–3807, 2021.

[6] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2015.

[7] O. Ronneberger, P. Fischer, and T. Brox, “U-net: convolutional networks for biomedical image segmentation,” in *Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241, Springer, Munich, Germany, 2015.

[8] C. G. B. Yogananda, B. R. Shah, M. Vejdani-Jahromi et al., “A fully automated deep learning network for brain tumor segmentation,” *Tomography*, vol. 6, no. 2, pp. 186–193, 2020.

[9] K. Kamnitsas, C. Ledig, V. F. J. Newcombe et al., “Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation,” *Medical Image Analysis*, vol. 36, pp. 61–78, 2017.

[10] K. He, X. Zhang, S. Ren et al., “Deep residual learning for image recognition,” in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, 2016.

[11] K. He, X. Zhang, S. Ren, and J. Sun, “Identity mappings in deep residual networks,” in *Proceedings of the 2016 European Conference on Computer Vision ECCV*, vol. 9980, pp. 630–645, Amsterdam, Netherlands, 2016.

[12] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[13] C. Szegedy, W. Liu, Y. Jia et al., “Going deeper with convolutions,” in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, Boston, MA, USA, 2015.

[14] F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” in *International Conference on Learning Representations (ICLR)*, San Juan, PR, USA, 2016.

[15] F. Hoseini, A. Shahbahrami, and P. Bayat, “An efficient implementation of deep convolutional neural networks for MRI segmentation,” *Journal of Digital Imaging*, vol. 31, no. 5, pp. 738–747, 2018.

[16] M. B. Naceur, R. Saouli, M. Akil, and R. Kachouri, “Fully automatic brain tumor segmentation using end-to-end incremental deep neural networks in MRI images,” *Computer Methods and Programs in Biomedicine*, vol. 166, pp. 39–49, 2018.

[17] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: a deep convolutional encoder-decoder architecture for image segmentation,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, pp. 2481–2495, 2017.

[18] M. Shaikh, G. Anand, G. Acharya, A. Amrutkar, V. Alex, and G. Krishnamurthi, “Brain tumor segmentation using dense fully convolutional neural network,” in *Proceedings of the 3rd International Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pp. 309–319, Springer, Granada, Spain, 2018.

[19] O. Oktay, J. Schlemper, L. L. Folgoc et al., “Attention U-net: learning where to look for the pancreas,” 2018, https://arxiv.org/abs/1804.03999.