GAU-Net: U-Net Based on Global Attention Mechanism for brain tumor segmentation

Xiuling Gan¹, Lidan Wang¹,3,4,5,*, Qi Chen¹, Yongjie Ge¹, Shukai Duan²,3,4,5

¹School of Electronic & Information Engineering, Southwest University, Chongqing 400715, China
²College of Artificial Intelligence, Southwest University, Chongqing 400715, China
³Brain-inspired Computing & Intelligent Control of Chongqing Key Lab, Chongqing 400715, China
⁴National & Local Joint Engineering Laboratory of Intelligent Transmission and Control Technology, Chongqing 400715, China
⁵Chongqing Brain Science Collaborative Innovation Center, Chongqing 400715, China

* Corresponding author’s e-mail: ldwang@swu.edu.cn

Abstract. Deep learning has shown great advantages in biomedical image segmentation. The classic model U-Net uses a stacked encoding-decoding structure of convolution operations for feature extraction and pixel-level classification. The stacking of convolutional layers can expand the receptive field, but it is still a local operation and cannot capture long-distance dependence. Therefore, in this work, we propose a Global Attention Mechanism that combines channel attention module and spatial attention module and integrates different convolutions in it. Besides, we design a residual module for the traditional up and down sampling blocks. And finally, we combine them with U-Net to propose a new global attention network GAU-Net. We perform experiments on the dataset BraTS2018. Our model has increased the mIoU from 0.65 to 0.75 with only 5.4% of U-Net parameters. At the same time, the inference time is also significantly shortened with relatively good performance.

1. Introduction

At present, image segmentation tasks have penetrated various fields, such as road scene segmentation involved in autonomous driving, lesion segmentation in medical images, and object positioning in satellite images. Since the introduction of AlexNet [1], many classic convolutional neural networks (CNNs) have been proposed in image segmentation tasks [2-8] and have achieved good results. However, the accuracy of these networks is still unsatisfactory and there is still much room for improvement. Many works [9,10] proposed that capturing long-distance dependence in segmentation tasks is crucial in deep learning networks, especially medical images that require long-distance dependence context information. However, the current deep learning models generally use CNN as the core, and the long-distance dependence obtained by the receptive fields stacked by the convolution kernel is still local. The Kaiming team proposed the non-local operations [9] in 2018 and pointed out that non-local operations can obtain the long-distance dependence of any two positions and are not limited by distance.
In image segmentation tasks, the problem of class imbalance is serious especially in medical image segmentation, so attention mechanism is also needed. A large number of segmentation networks based on the attention mechanism [11-14] have also been proposed. Therefore, this paper integrates non-local operations and proposes a GAU-Net for brain tumor segmentation, so the main contributions of this paper are fourfold: (1) we propose a Global Attention Mechanism that combines spatial attention module and channel attention module. Its essence is still a non-local operation used to capture long-distance dependencies and solve the problem that convolution operations can only extract local information. (2) Different convolutions are used in the Global Attention Mechanism module to calculate the correlation between two positions, that is, to improve the diversity and characterization capabilities of the extracted features. (3) We use residual connections in convolution, sampling, and GAM modules. That is, short-distance skip connections to obtain multi-scale features and improve the segmentation effect of the network. (4) Combine the proposed GAM module with U-Net to make the model more focused on the key parts to be segmented, and achieve high segmentation accuracy with very few parameters.

2. GAU-Net

2.1. Overall network architecture

The network structure of Global Attention U-Net is shown in Fig. 1. The network mainly includes input block, down block, bottom block, up block, and output block. The combination of residual connection, Fire Module, and global attention mechanism effectively compress the model scale, while ensuring even improving the segmentation ability of the network. We have added GAM between the up and down sampling paths to enhance the long-distance dependence of the feature maps generated in the subsequent decoding process and encoding process.

![Fig.1 Overall architecture of the global attention U-Net](image)

2.2. Residual Blocks

We design the convolutional layer and sampling layer as a residual structure, which is the down block and up block of the GAU-Net. We use the Fire Module proposed in [15] to implement the convolution operation. As shown in Fig. 2(a), the convolution method of this module can greatly reduce the number of model parameters, which include squeeze layer and expand layer. Moreover, we add the Batch Normalization (BN) layer and the ReLU activation function after each convolutional layer. Finally, concatenate the feature maps extracted by the two different convolution kernels of the expand layer to obtain the convolved feature maps.

A large number of residual structures are used in our network design. The use of residual structures can strengthen the fusion of multi-scale features without increasing the number of parameters. In order to balance the number of parameters and the calculation flow, we design the convolutional layer and
the sampling layer in a residual block. In the down block, we use the Fire Module for feature extraction in the main path. Then use the maximum pooling layer to downsample the feature maps. And in the bypass connection, we use a convolutional layer with the stride of 2 to extract and sample features at the same time to reduce feature loss. The up block and the down block are corresponding, the difference is that the transposed convolution is used for up-sampling in the bypass connection. The details of our proposed up and down blocks are shown in Fig. 2(b) and Fig. 2(c).

2.3. Global Attention Mechanism

In order to fuse global features so that the extracted features depend on all positions in the input feature map and can be expressed in space and channels, we propose a Global Attention Mechanism (GAM) that captures long-distance dependence.

The GAM we proposed contains two modules: the spatial attention module and the Channel attention module. The definition of non-local operations proposed in [9] is shown in formula n,

\[ y_i = \frac{1}{C(x)} \sum_j f(x_i, x_j)g(x) \]  

(1)

where \( \frac{1}{C(x)} \sum_j f(x_i, x_j) \) is a normalized function, \( f(x_i, x_j) \) is used to calculate the correlation between position \( i \) and all other positions \( j \), and \( g(x) \) is a unary function used to transform information. Among
them, $f(x_i, x_j)$ can be realized by Gaussian function, embedded Gaussian function, dot multiplication, and other operations.

In our paper, we consider $f(x_i, x_j)$ is the embedded Gaussian function:

$$f(x_i, x_j) = e^{\theta \phi(x_i, x_j)}$$  \hspace{1cm} (2)

where $\theta(x_i) = W_x \phi(x_i)$, $\phi(x_j)$ are two embeddings, which is manifested in our module by using convolution operation to realize embedding. We use different convolutions in the three paths of Query ($Q$), Key ($K$), and Value ($V$) in the two attention modules to extract richer feature maps, so we use $Y_s$ and $Y_c$ to represent the output of the Spatial attention module and Channel attention module respectively, which can be expressed as follows:

$$Y_s = \text{softmax}(W_Q \cdot X' \cdot \phi(x_i)W_k)^T W(x)$$  \hspace{1cm} (3)

$$Y_c = \text{softmax}(X' \cdot W_k^T \cdot W_Q \cdot x)V(x)$$  \hspace{1cm} (4)

where $x'$ represents the output of the 1×1 convolutional layer.

In GAM, we also use the residual connection. We propose two connection forms, parallel and cascade, and the effect is verified by experiments in Chapter 4. In addition, we compress the input channel $C$ of the convolutional layer to reduce the parameters. We set the number of channels to 1/2 of the input in the spatial attention module and channel attention module. The number of input and output channels of the GAM can be ensured to be consistent, which is convenient for residual connection. And it is also conducive to embedding the module into different networks. The GAM of the two connection modes is shown in Fig. 5.

**Fig. 5 Connection modes of channel attention module and spatial attention module**

### 3. Experimental and Results

In this section, we introduce some common settings in experiments, perform detailed experiments on the BraTS2018 dataset, and use the U-Net proposed in [3] as the baseline model.

**Datasets:** The BraTS2018 dataset contains MRI images of 351 brain tumor patients, which segment the whole tumor (WT), tumor core (TC), and enhanced tumor (ET). The training set contains 285 samples with manual labels. There are 66 samples unlabelled in the test set. Each sample contains 155 MRI images with four modalities and size of 240×240, namely T1, T1ce, T2 and Flair. To facilitate the verification of the experimental results, we divide the images with ground truth into the training set and the test set at a ratio of 9:1 and eliminate the slices without lesions. The specific data quantity used in the experiment is shown in Table 1.

| Dataset | cases | All images | After removing |
|---------|-------|------------|----------------|
| BraTS2018 | Training set | 255 | 39525 | 17031 |
|         | Testing set  | 30  | 4650  | 1893  |

Table 1. The number of data used in experiments.
Overall 285 44175 18924

Evaluation metric: In order to quantitatively evaluate the performance of our proposed model, pixel intersection-over-union (IoU), average pixel intersection-over-union (mIoU), Dice similarity coefficient (Dice), sensitivity and positive predictive value (PPV).

Training: We use Pytorch to implement our model. All experiments are realized on an NVIDIA GTX 1080 TI 11GB GPU. We set the initial learning rate to $3 \times 10^{-4}$ and set the batch size to 30 according to the allowed capacity of the GPU. Adam is used for training and its parameters are default values. The ratio of the number of convolution kernels between the squeeze layer and the expand layer in the Fire Module is 0.25. The number of training epochs is set to 1000, and the early stopping method is used to prevent network overfitting.

3.1. Comparison with Baseline

Quantitative Evaluations. In Table 2, we report the performance of our network on the data set BraTS2018. Compared with the baseline model U-Net, our model achieves very good performance and is significantly better than U-Net. Fig. 6 is the curves of the model training process.

| models     | IoU WT | Dice WT | Sensitivity WT | PPV WT | IoU TC | Dice TC | Sensitivity TC | PPV TC | IoU ET | Dice ET | Sensitivity ET | PPV ET |
|------------|--------|---------|----------------|--------|--------|---------|----------------|--------|--------|---------|----------------|--------|
| Baseline   | 0.68   | 0.70    | 0.58           | 0.82   | 0.82   | 0.76    | 0.84           | 0.79   | 0.86   | 0.84    | 0.79           |        |
| GAU-Net    | 0.76   | 0.81    | 0.67           | 0.82   | 0.85   | 0.76    | 0.84           | 0.79   | 0.88   | 0.89    | 0.82           |        |

Fig. 6 Training curves of the model

Fig. 7 Images of models in prediction on BraTS2018
Qualitative Evaluations. The qualitative evaluation on the BraTS2018 dataset is shown in Figure n. The first two rows are the original images displayed in the open-source software ITK-SNAP and manual segmentation results. The following are the segmentation results of U-Net and our proposed GAU-Net, as shown in Fig. 7. Our model is proved to be more advantageous in the segmentation of tumor edges and details. In addition, because the long-distance dependence and detailed features are effectively captured, the performance in pixel classification is also very good. Compared with U-Net, there is almost no misclassification of pixels into other categories.

Model Size Analysis. We report the size of U-Net in Table 3. Compared with our model, our model has fewer parameters, requiring only 5.4% of the U-Net parameters, but achieves better results and shortens the inference time. Therefore, it is effectively proved that our proposed GAM module improves the performance of the model by capturing the global information of the image.

| models          | Number of parameters size | mIoU | Inference |
|-----------------|---------------------------|------|-----------|
| U-Net           | 3940067                   | 150.30 | 0.65   | 6.17±0.5 |
| GAU-Net         | 214603                    | 8.19  | 0.75   | 4.42±0.5 |

3.2. Ablation Studies of Different Modules

Parallel v.s. Cascade Connection GAM. Our proposed GAM combines two modules of spatial attention module and channel attention module. The connection of these two modules will also directly affect the effectiveness of the model. Therefore, we have done comparative experiments on the internal structure of GAM. We conducted experiments on the parallel connection and cascade connection of two modules on the BraTS2018. And we compared the sizes of the two modules based on the input size of 20×20×256. The performances are shown in Table 4.

| models                              | Number of parameters size | mIoU of model |
|-------------------------------------|---------------------------|--------------|
| GAU-Net with Parallel GAM           | 1081856                   | 4.13         |
| GAU-Net with Cascade GAM            | 692864                    | 2.64         |

Experiments have proved that the cascading spatial attention and channel attention in GAM is more effective. The cascaded GAM takes the input as a whole. The input first passes through the spatial attention module, paying more attention to the position of the target in the image, and then passes through the channel attention module to focus on what the target is. After such a global attention module, more effective features for segmentation targets are extracted. Therefore, we use the cascaded GAM in the final proposed network GAU-Net.

Number and Position of GAM. In addition to the influence of the internal structure of the attention module on network performance, the location and number of GAMs when combined with the network will also affect network performance. Therefore, we conducted ablation experiments on this. The GAU-Net we proposed is a completely symmetrical up and downsampling path, including three up and down-sampling modules. We can connect up to three GAM modules in the network. From top to bottom, we record the three down-sampling connected GAM modules as 1, 2, and 3 respectively. Table 5 reports the number and location of our connected GAMs and the data set BraTS2018 effect.

| GAM | Dice  | mIoU of model |
|-----|-------|---------------|
|     | WT    | TC            | ET            | model |
| 3   | 0.8015 | 0.8363       | 0.7374       | 0.7249 |
| 1, 3| 0.7949 | 0.8237       | 0.7239       | 0.7152 |
2, 3  0.8043  0.8401  0.7396  0.7322
1, 2, 3 (Our model)  0.8261  0.8545  0.7616  0.75

4. Conclusion
In this paper, we propose a Global Attention U-Net for brain tumor segmentation by capturing long-distance dependencies to improve network performance. By extending non-local operators, we propose a Global Attention Mechanism that combines channel attention module and spatial attention module and incorporates different convolutions into it. To ensure network performance, we also use residual connections for up and down sampling blocks and attention modules. We connect the GAM module to the branch of the skip connection to strengthen the combination of low-level texture information and high-level semantic information. By capturing detailed information and long-distance dependence, our network has achieved good performance with few parameters. However, our network architecture still has a lot of room for improvement, and its segmentation effect is still far from clinical use. So the improvement of network performance is worthy of further discussion and research.

Acknowledgments
The work was supported by the National Key R&D Program of China (Grant No. 2018YFB1306600), the National Natural Science Foundation of China (Grant Nos.62076207, 62076208, U20A20227).

References
[1] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM. 2012, 60(6): 84-90.
[2] Shelhamer E, Long J, Darrell T. Fully Convolutional Networks for Semantic Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2016, 39(4): 640-651.
[3] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]. Munich, Germany: Springer Verlag, 2015.
[4] Chen L, Papandreou G, et al. Semantic image segmentation with deep convolutional nets and fully connected CRFs[C]. San Diego, CA, United states: International Conference on Learning Representations, ICLR, 2015.
[5] Chen L C, Papandreou G, et al. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs[J]. IEEE Trans Pattern Anal Mach Intell. 2018, 4(40): 834-848.
[6] Chen L, Papandreou G, et al. Rethinking Atrous Convolution for Semantic Image Segmentation[J]. 2017.
[7] Chen L, Zhu Y, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]. Munich, Germany: Springer Verlag, 2018.
[8] Milletari F, Navab N, Ahmadi S. V-Net: Fully convolutional neural networks for volumetric medical image segmentation[C]. Stanford, CA, United states: Institute of Electrical and Electronics Engineers Inc., 2016.
[9] Wang X, Girshick R, et al. Non-local Neural Networks[C]. Salt Lake City, UT, United states: 2017.
[10] Zhao H, Shi J, et al. Pyramid scene parsing network[C]. Honolulu, HI, United states: Institute of Electrical and Electronics Engineers Inc., 2017.
[11] Fu J, Liu J, Tian H, et al. Dual attention network for scene segmentation[C]. Long Beach, CA, United states: IEEE Computer Society, 2019.
[12] Yu Q, Xia Y, et al. Thickened 2D Networks for Efficient 3D Medical Image Segmentation[J]. 2019.
[13] Zhu Z, Xu M, et al. Asymmetric non-local neural networks for semantic segmentation[C]. Seoul, Korea, Republic of: Institute of Electrical and Electronics Engineers Inc., 2019.
[14] Chen L, Yang Y, et al. Attention to Scale: Scale-Aware Semantic Image Segmentation[C].
[15] Iandola F, Han S, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size[J]. 2016.