doi:10.3772/j.issn.1006-6748.2022.03.003

Completed attention convolutional neural network for MRI image segmentation[①]

ZHANG Zhong(张　重)[②*], LV Shijie[*], LIU Shuang[*], XIAO Baihua[**]

([*] Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission,
Tianjin Normal University, Tianjin 300387, P. R. China)

([**] State Key Laboratory for Management and Control of Complex Systems, Institute of Automation,
Chinese Academy of Sciences, Beijing 100190, P. R. China)

Abstract

Attention mechanism combined with convolutional neural network (CNN) achieves promising performance for magnetic resonance imaging (MRI) image segmentation, however these methods only learn attention weights from single scale, resulting in incomplete attention learning. A novel method named completed attention convolutional neural network (CACNN) is proposed for MRI image segmentation. Specifically, the channel-wise attention block (CWAB) and the pixel-wise attention block (PWAB) are designed to learn attention weights from the aspects of channel and pixel levels. As a result, completed attention weights are obtained, which is beneficial to discriminative feature learning. The method is verified on two widely used datasets (HVSMR and MRBrainS), and the experimental results demonstrate that the proposed method achieves better results than the state-of-the-art methods.

Key words: magnetic resonance imaging (MRI) image segmentation, completed attention convolutional neural network (CACNN)

0 Introduction

Magnetic resonance imaging (MRI) is one of the fundamental technologies to detect different diseases, such as brain tumour, cardiovascular lesions and spinal deformity[1]. This technology is widely used due to its non-invasive characteristic and multi-modality information. As the basis of medical image technologies, MRI medical image segmentation is valuable for research and practical application. For example, it can assist doctors in clinical diagnosis, surgical guidance and so on.

With the development of deep learning, convolutional neural network (CNN) dominates the field of MRI image segmentation. Some methods[2-3] apply the fully convolutional network (FCN) for segmentation, which changes the fully connected layer into the convolution layer and fuses the features of pooling layers and the last convolution layer. Ronneberger et al.[4] designed the U-shape architecture network (U-Net) for biomedical image segmentation, which utilizes the contracting path and the symmetric expanding path along with skip connections to obtain the pixel-wise prediction. Because of the superiority of U-Net, some variants[5-7] were proposed to apply in the field of medical image segmentation.

Recently, the attention mechanism[8], which is prone to pay attention to the important parts of an image rather than the whole one, is introduced into the medical image segmentation. With the attention mechanism, the attention-aware features were generated to adaptively adjust the weights of features[9]. Pei et al.[10] proposed the position attention module and the channel attention module in single scale so as to make the network concentrate on the core location of colorectal tumour. Mou et al.[11] presented CS2-Net which applies the self-attention mechanism to learn rich hierarchical features for medical image segmentation. However, the above-mentioned attention-based methods only learn attention weights from single scale, which is difficult to obtain completed attention information.

In this paper, a novel method named completed attention convolutional neural network (CACNN) is proposed for MRI image segmentation, which learns

[①] Supported by the National Natural Science Foundation of China (No. 62171321), Tianjin Municipal Natural Science Foundation (No. 20JCZDJC00180, 19JCZDJC31500) and the Open Projects Program of National Laboratory of Pattern Recognition (No. 202000002).

[②] To whom correspondence should be addressed. E-mail: s.liu@tjnu.edu.cn.

Received on Sep. 17, 2021

channel-wise attention weights and pixel-wise attention weights simultaneously. To this end, CACNN is designed as the symmetric structure including the encoder, the decoder, the channel-wise attention block (CWAB), and the pixel-wise attention block (PWAB). Specifically, CWAB learns the attention weights for each channel so as to adaptively fuse the feature maps of the encoder and the decoder at the same scale. Meanwhile, PWAB learns the attention weights for each pixel in order to fuse the feature maps from different blocks of decoder. In a word, CACNN could learn the attention weights for different aspects, which forces the deep network to focus on extracting the discriminative features for MRI image segmentation.

1 Approach

The framework of the proposed CACNN is shown in Fig. 1, which includes the encoder, the decoder, CWAB, and PWAB. In this section, each component of CACNN is introduced in detail.

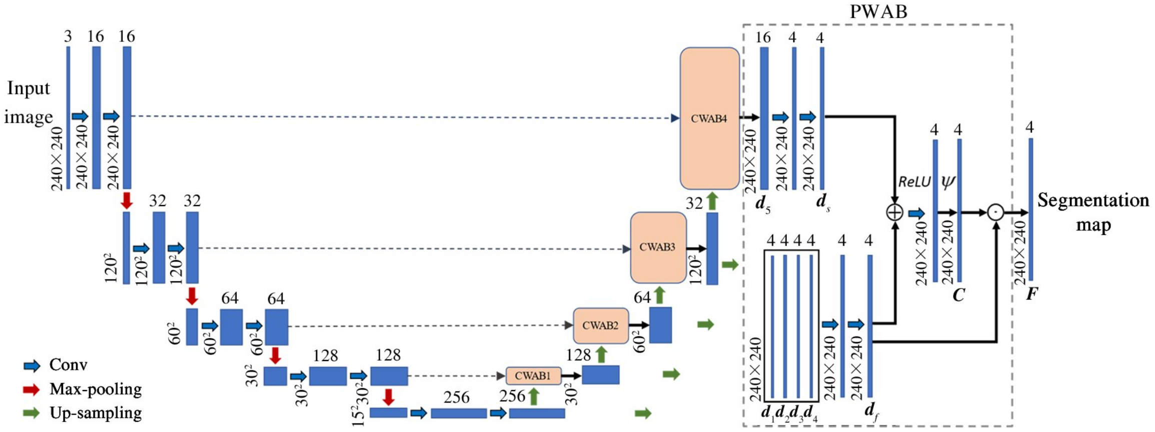

Fig. 1 The framework of the proposed CACNN

1.1 The structure of CACNN

CACNN is the symmetric structure where the encoder and the decoder both contain four blocks. As for the encoder, each block consists of two convolution layers and one max-pooling layer, where the kernel size of convolution layer is 3 × 3 with the sliding stride of 1 and the kernel size of max-pooling layer is 2 × 2 with the sliding stride of 2. As for the decoder, each block includes the up-sampling operations to obtain the feature maps with the same size of the corresponding encoder block. Instead of the skip connections, the feature maps of the corresponding blocks from the encoder and the decoder are fed into CWAB. Afterwards, the outputs of CWAB1-3 and the minimum scale feature maps are fed into PWAB after up-sampling operations. Meanwhile, the output of CWAB4 is also as the input of PWAB to generate the final segmentation map.

1.2 Channel-wise attention block

In order to fuse the information from the encoder, some traditional segmentation methods[2-4] adopt the skip connections to directly concatenate the feature maps from the encoder and the decoder. However, the skip connections neglect the importance of different channels of feature maps. Hence, CWAB is proposed to assign different attention weights to each channel. Since there are four blocks of the encoder and the decoder, four CWABs are inserted into CACNN. The structures of four CWABs are similar, and therefore taken CWAB1 is an example. The structure of CWAB1 is shown in Fig. 2. The feature maps ($\boldsymbol{I}$ and $\boldsymbol{T}$) of the corresponding block from the encoder and the decoder are as the input of CWAB1, and they are first conducted by the max-pooling operation to obtain $\boldsymbol{H} \in R^{1\times128}$ and $\boldsymbol{Q} \in R^{1\times128}$, respectively. Then, the attention weights can be obtained as

$$\boldsymbol{W} = \mathrm{softmax}(\theta(\boldsymbol{A} \times [\boldsymbol{H} \parallel \boldsymbol{Q}])) \qquad (1)$$

where, $\parallel$ denotes the concatenation operation, $\boldsymbol{A}$ is the learnable transformation vector; × indicates the matrix multiplication operation; θ is the non-linear transformation, and it is implemented by the LeakReLU activation function in the experiment. As a result, $\boldsymbol{W}$

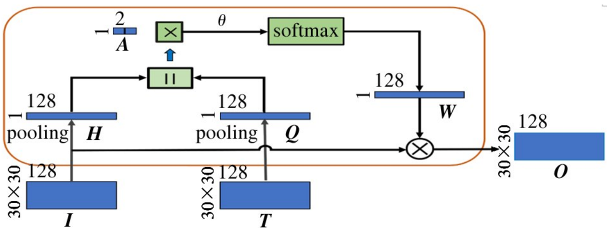

Fig. 2 The structure of CWAB

contains the attention weights for each channel, and the output of CWAB is represented as

$$\boldsymbol{O} = \boldsymbol{W} \otimes \boldsymbol{I} \tag{2}$$

where $\otimes$ indicates the channel-wise multiplication. In a word, the output of CWAB reflects the important of feature maps, and therefore the representation ability is improved.

1.3 Pixel-wise attention block

Most existing segmentation methods[4-5] utilized the feature maps of the last block from the decoder to calculate the final segmentation map, and they only learn the feature maps from single scale. Therefore, PWAB is proposed to fuse feature maps from different scales in order to obtain accurate segmentation map as shown in the right part of Fig. 1. Firstly, the feature maps from different blocks are conducted by the up-sampling operations to obtain the feature maps with the same size, i. e., 4 × 240 × 240. Then, the feature maps are concatenated in a weighted way and utilize the convolution operation to obtain $\boldsymbol{d}_f \in R^{4\times240\times240}$:

$$\boldsymbol{d}_f = \phi[4\boldsymbol{d}_1 \parallel 3\boldsymbol{d}_2 \parallel 2\boldsymbol{d}_3 \parallel \boldsymbol{d}_4] \tag{3}$$

where ϕ represents the convolution operation. The output of CWAB4 is $\boldsymbol{d}_5 \in R^{16\times240\times240}$, and then $\boldsymbol{d}_s \in R^{4\times240\times240}$ is obtained after convolution operations. The attention weights are defined as

$$\boldsymbol{C} = \psi\{\phi(\mathrm{ReLU}(\boldsymbol{d}_f + \boldsymbol{d}_s))\} \tag{4}$$

where ψ represents the Sigmoid function. The final segmentation map $\boldsymbol{F}$ is formulated as

$$\boldsymbol{F} = \boldsymbol{C} \odot \boldsymbol{d}_f \tag{5}$$

where $\odot$ indicates the pixel-wise multiplication. As a result, it could fuse the feature maps of multiple scales by assigning attention weights to each pixel. In order to train the network, the cross-entropy function is employed as the loss.

2 Experiments and analysis

2.1 Datasets and implementation details

The experiments is conducted on two challenging datasets: HVSMR[12] dataset and MRBrainS[13] dateset. HVSMR has one modality information, i. e., T2 sequences and it aims to segment blood pool and myocardium in cardiovascular MR images. It includes 1868 training slices and 1473 test slices. The images in HVSMR have different size, and the size of the original images is maintained and fed into the network. MRBrainS contains MR brain scans of three modality information, i. e., T1, T1 inversion recovery and FLAIR sequences. The task of MRBrainS is to segment cerebrospinal fluid, gray matter and white matter. It consists of 104 slices for training and 70 slices for testing, where the size of each slice is 240 × 240. The training images are conducted by a skull stripping pre-processing.

The Adam algorithm is adopted for deep network optimization with the weight decay of 5×10^{-4} and the learning rate of 6×10^{-4}. Furthermore, the batch normalization is utilized in CWABs.

In order to evaluate the performance of MRI image segmentation, two matrices is employed, i. e., pixel accuracy and dice score[1,10]. The pixel accuracy indicates the ratio between the number of correctly classified pixels and the total number of pixels. The dice score reflects the overlap between the prediction results and the ground-truth. The two matrices are defined as

$$Acc = (TN + TP)/(TP + TN + FP + FN) \tag{6}$$

$$Dice = 2 \times TP/(FP + 2 \times TP + FN) \tag{7}$$

where TP is the true positive, FP is the false negative, TN is the true negative and FN is the false negative.

2.2 Comparing to the state-of-the-art methods

The proposed CACNN is compared with the state-of-the-art methods, such as FCN[2], SegNet[14] and U-Net[4] where different encoders are utilized, i. e. VGG16, ResNet50 and U-Net structure. The pixel accuracy and the dice score are listed in Table 1, where the following four conclusions can be drawn. Firstly, CACNN gains the best results on the two datasets, because it learns completed attention weights including the channel level and pixel level. The proposed CACNN achieves the pixel accuracy and the dice score of 94.15% and 88.39% on the HVSMR dataset, and 97.13% and 90.48% on the MRBrainS dataset. Secondly, the proposed two attention mechanisms (CWAB and PWAB) both boost the segmentation performance compared with the baseline (U-Net). Note that the backbone of CACNN is designed as U-Net, and therefore it is reasonable to treat U-Net as the baseline. Compared with U-Net, CWAB and PWAB raise the dice score by 0.12% and 1.42% on the HVSMR dataset. It proves that the channel-level attention and the pixel-level attention are both essential for performance improvement. Thirdly, U-Net structure performs better than other network architecture. For example, U-Net with VGG 16 and U-Net with ResNet 50 obtain the best performance for the same encoder (VGG 16 or ResNet 50). Hence, U-Net is chosen as the backbone of CACNN. Finally, MRBrainS contains three kinds of modality information, while HVSMR includes one kind. Comparing their performance, it shows that multiple information is beneficial to the performance improvement.

Table 1 Comparison on the HVSMR dataset and the MRBrainS dataset with different methods

Model	HVSMR		MRBrainS	
	Pixel ***Acc***	***Dice***	Pixel ***Acc***	***Dice***
FCN with VGG 16	0.9165	0.8368	0.957	0.8637
SegNet with VGG 16	0.8928	0.7118	0.9484	0.8294
U-Net with VGG 16	0.9109	0.8201	0.9696	0.8991
FCN with ResNet 50	0.9095	0.8266	0.9488	0.8347
U-Net with ResNet 50	0.9192	0.8371	0.9701	0.9039
FCN with U-Net	0.9179	0.8295	0.9564	0.8618
SegNet with U-Net	0.9027	0.8099	0.9506	0.8448
Baseline (U-Net)	0.9270	0.8593	0.9705	0.9021
CWAB	0.9287	0.8605	0.9707	0.9024
PWAB	0.9368	0.8735	0.9710	0.9038
CACNN	0.9415	0.8839	0.9713	0.9048

Fig. 3 illustrates some results of CACNN and other methods on the two datasets. It shows that CACNN is the superiority on dealing with detail information because of the completed attention learning.

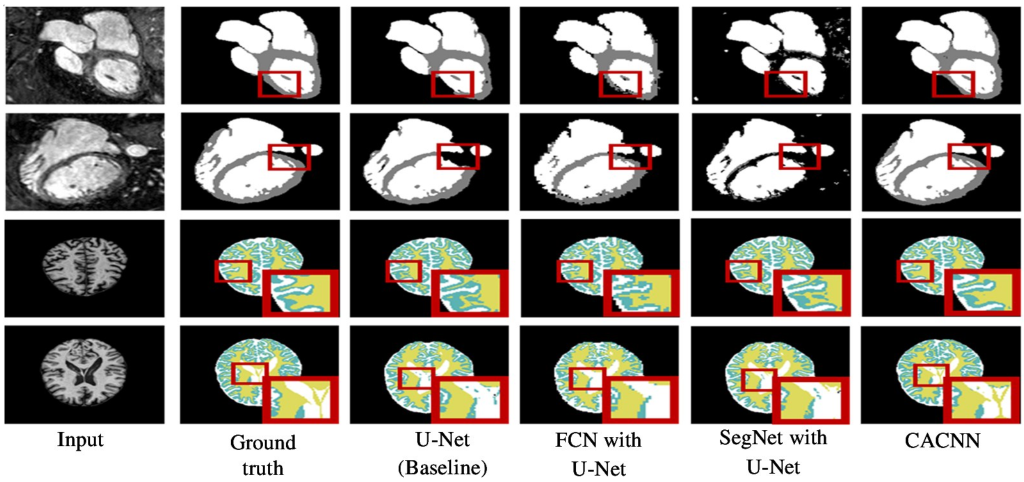

Fig. 3 Some results of CACNN and other methods (The first two rows are the samples from HVSMR and the last two rows are samples from MRBrainS)

2.3 Parameter analysis

There are three key parameters in CACNN and therefore a series of experiments are conducted to search the optimal parameter values. Firstly, it tests different number of fused feature maps for PWAB in Eq. (3), and the results are listed in Table 2. It shows the performance improvement with the number of fused feature maps. Hence, the feature maps are fused from four blocks. Then, it study which feature maps to multiply with the attention weights for CWAB and PWAB respectively in Eq. (2) and Eq. (5). From Table 3 and Table 4, it shows that $\boldsymbol{I}$ and $\boldsymbol{d}_f$ are optimal in CACNN.

Table 2 Comparison of different numbers of fused feature maps on the MRBrainS dataset

Different feature maps	Pixel ***Acc***	***Dice***
$\boldsymbol{d}_1$	0.9615	0.8611
$\boldsymbol{d}_1\boldsymbol{d}_2$	0.9623	0.8624
$\boldsymbol{d}_1\boldsymbol{d}_2\boldsymbol{d}_3$	0.9678	0.8935
$\boldsymbol{d}_1\boldsymbol{d}_2\boldsymbol{d}_3\boldsymbol{d}_4$	0.9713	0.9048

Table 3 Comparison of different feature map combination in CWAB on the MRBrainS dataset

Feature maps	Pixel ***Acc***	***Dice***
$\boldsymbol{T}$ with $\boldsymbol{W}$	0.9706	0.9010
$\boldsymbol{I}$ with $\boldsymbol{W}$	0.9713	0.9048

Table 4 Comparison of different feature map combination in PWAB on the MRBrainS dataset

Feature maps	Pixel ***Acc***	***Dice***
$\boldsymbol{d}_s$ with $\boldsymbol{C}$	0.9686	0.8955
$\boldsymbol{d}_f$ with $\boldsymbol{C}$	0.9713	0.9048

3 Conclusion

CACNN is proposed for MRI image segmentation to learn completed attention weights. CACNN mainly contains CWAB and PWAB which learn attention weights for each channel and each pixel, respectively. With the completed attention weights, the deep network focuses on extracting the discriminative features. CACNN is validated on two datasets HVSMR and MRBrainS, and the experimental results demonstrate that the proposed method outperforms the other methods.

References

[1] WEN Y, XIE K, HE L. Segmenting medical MRI via recurrent decoding cell[C]//AAAI Conference on Artificial Intelligence, New York, USA, 2020:12452-12459

[2] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]// IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 3431-3440

[3] BEN-COHEN A, DIAMANT I, KLANG E, et al. Fully convolutional network for liver segmentation and lesions detection[C]// Proceedings of the 1st Workshop on Large-scale Annotation of Biomedical Data and Expert Label Synthesis, Athens, Greece, 2016: 77-85

[4] RONNEBERGER O, FISCHER P, BROX T. U-Net: convolutional networks for biomedical image segmentation[C]//The 18th International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 2015:234-241

[5] CICEK O, ABDULKADIR A, LIENKAMP S, et al. 3D U-Net: learning dense volumetric segmentation from sparse annotation[C]// The 19th International Conference on Medical Image Computing and Computer-Assisted Intervention, Athens, Greece, 2016:424-432

[6] ZHANG J, JIN Y, XU J, et al. MDU-Net: multi-scale densely connected U-Net for biomedical image segmentation[EB/OL]. http://arxiv.org/abs/1812.00352:arXiv, (2018-12-02),[2021-09-17]

[7] JAFARI M, AUER D, FRANCIS S, et al. DRUNet: an efficient deep convolutional neural network for medical image segmentation[C]//IEEE International Symposium on Biomedical Imaging, Iowa City, USA, 2020: 1144-1148

[8] XIE Y, LIANG R, LIANG Z, et al. Attention-based dense LSTM for speech emotion recognition[J]. *IEICE Transactions on Information and Systems*, 2019, 102(7): 1426-1429

[9] VELICKOVIC P, CUCURULL G, CASANOVA A, et al. Graph attention networks[EB/OL]. http://arXiv.org/abs/1710.10903:arXiv, (2017-10-30),[2021-09-17]

[10] PEI Y, MU L, FU Y, et al. Colorectal tumor segmentation of CT scans based on a convolutional neural network with an attention mechanism[J]. *IEEE Access*, 2020, 8: 64131-64138

[11] MOU L, ZHAO Y, FU H, et al. CS2-Net: Deep learning segmentation of curvilinear structures in medical imaging[J]. *Medical Image Analysis*, 2021, 67: 101874

[12] PACE D F, DALCA A V, GEVA T, et al. Interactive whole-heart segmentation in congenital heart disease[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 2015: 80-88

[13] MENDRIK A M, VINCKEN K L, KUIJF H J, et al. MRBrainS challenge: online evaluation framework for brain image segmentation in 3T MRI scans[J]. *Computational Intelligence and Neuroscience*, 2015, 2015:B13696

[14] BADRINARAYANAN V, KENDALL A, CIPOLLA R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, 39(12): 2481-2495

ZHANG Zhong, born in 1986. He is a professor at Tianjin Normal University, China. He is a senior member of IEEE. He received the Ph. D degree from Institute of Automation, Chinese Academy of Sciences, Beijing, China. He has published about 110 papers in international journals and conferences such as IEEE TFS, PR, IEEE TCSVT, IEEE TIFS, Signal Processing (Elsevier), CVPR, ICPR and ICIP. His research interests include computer vision, remote sensing and deep learning.