MixNet: Multi-modality mix network for brain segmentation

Abstract

In this work we propose a new 2D convolutional neural network, called MixNet, to tackle the brain tissue segmentation problem in multi-modal MR images (including T1, T1-IR and T2-FLAIR). The network is organized as 3 relatively separate flows, each of which is responsible for processing one modality. Information of different flows is exchanged once after a certain number of convolutional layers. Based on our experiments, mixing the information periodically performs better than handling several modalities completely independent or always together. Feature maps from different flows and different intermediate convolutional layers are combined to form the final feature map, aggregating both multi-modality and multi-scale information. We use the residual learning block [1] as the basic unit to avoid the degradation problem when training very deep network. Moreover, the dilated convolution is used to keep the resolution.

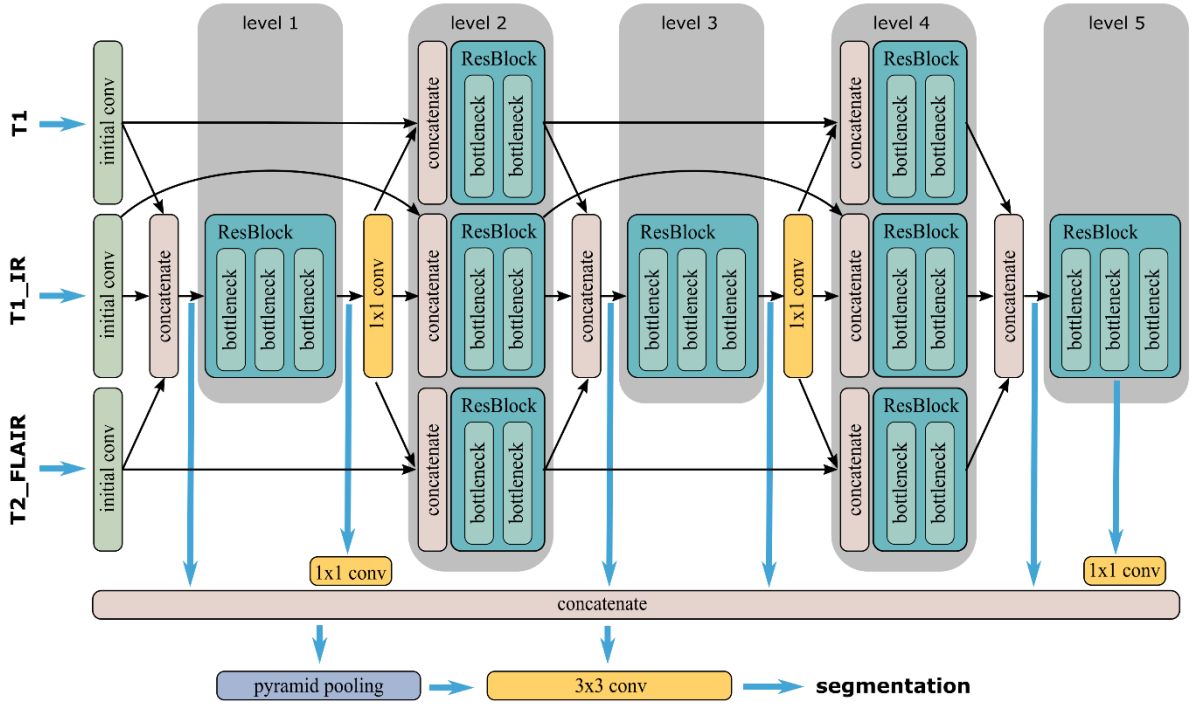

Figure 1: Structure of the proposed MixNet. Level 1, level 3 and level 5 play the role of periodic information summarization.

1. Structure of the network

There are many ways to use multi-modal images. Several predictors can be trained independently and each of them tries to segment the target tissues based on only one modality. At inference time, independent predictions made by these predictors will be merged with a certain fusion strategy. Another option is concatenating all modalities together as a multi-channel image.

Based on out experiments, neither way is optimal. The proposed MixNet summarizes information from different modality flows periodically. As shown in Fig. 1, level 1, level 3 and level 5 play such a role, the summarization is then fed back to each modality flow. As for the final feature map, feature maps of intermediate layers are aggregated to inject multi-scale information. Additionally, inspired by the success of PSPNet [3], we use a pyramid pooling module at the end of the network for global prior construction on the final feature map.

The ResBlock is the basic component of the network, which is composed of several bottleneck modules connected in series. As shown in Fig. 2, the bottleneck module takes a deep residual learning structure proposed by [1,2]. The main difference is that we use dilated convolution in the second convolutional layer. The dilated convolution avoids the resolution loss caused by max pooling, thus keeps more localization information in the segmentation task.

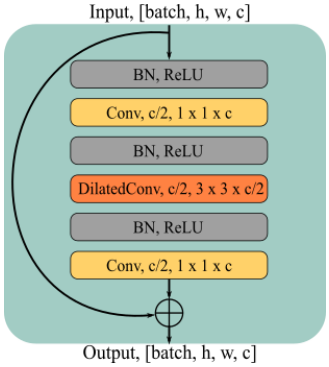

Figure2: Structure of the bottleneck module

At last, feature maps from lower layers has fewer features compared with higher layers, for example, the output of level 3 is twice that of level 1. We insert 1x1 convolution layers to maintain the balance of concatenated feature maps.

2. Training and prediction

2.1. Training

The network is trained with gradient descent optimization algorithm with Nesterov momentum. The momentum is set to 0.99. The initial learning rate is 2e-4 and is halved after each preset boundary epoch, which is 0.2, 0.4, 0.6, 0.75, 0.8, 0.85, 0.9 and 0.95 of the total number of training epochs.

The MRBrainS 2018 dataset is employed for training. Since the proposed network works on 2D images, the 3D volume is processed layer by layer from any of the three directions (the horizontal plane, the sagittal plane and the coronal plane). However, we train a main predictor with images of horizontal plane, since they process a higher resolution. Two additional predictors of the sagittal plane and the coronal plane are used for boosting.

2.2. Data augmentation

Our network is 45 layers deep, for which the MRBrainS 2018 training data is not particularly adequate. Thus, the data is heavily augmented with elastic deformation [4], scaling, rotation and translation. As for the sagittal plane and the coronal plane, the resolution in the horizontal and vertical directions are four times different. We only apply flipping, scaling and translation, which keeps horizontal lines horizontal and vertical lines vertical.

It is worth mention that excessive elastic deformation and scaling may lead to an unstable training. We use the scaling factors 0.9, 0.95, 1.05 and 1.1, elastic deformation factor $\alpha = 10$ and $\sigma = 4$ [4] in this work.

2.3. Prediction

As mentioned before, three predictors are trained with respect to three observation directions. Due to the higher resolution, the main predictor (on the horizontal plane) performs evidently better. But results from the other two predictors can be used for boosting. We compute the weighted sum of the 3 probability maps with the weight for the main predictor four times as large as the other two.

[1] K. He, X. Zhang, S. Ren and J. Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016.
[2] K. He, X. Zhang, S. Ren and J. Sun. Identity Mapping in Deep Residual Networks. In *ECCV*, 2016.
[3] H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia. Pyramid Scene Parsing Network. In *CVPR*, 2017.
[4] P. Y. Simard, D. Steinkraus and J. C. Platt. Best Practices for CNN Applied to Visual Document Analysis. In *ICDAR*, 2003.