3D Patchwise U-Net with Transition Layers for MR Brain Segmentation

Miguel Luna[1], Sang Hyun Park[1]

[1] Department of Robotics Engineering,
Daegu Gyeongbuk Institute of Science and Technology, Republic of Korea,
shpark13135@dgist.ac.kr

Abstract. To address the segmentation of gray matter, white matter, cerebrospinal fluid and other structures on brain Magnetic Resonance (MR) images, we propose a new 3D convolutional neural network that can utilize features from multi-modality images, *i.e.*, IR, T1 and FLAIR images.

1 Method

The proposed network consists of 3 stages, *i.e.*, encoder, transition, and decoder. The encoder takes the 3D input patches extracted from the set of FLAIR, T1 and IR images. The patch size was heuristically defined as $8 \times 24 \times 24$ voxels on the z, y and x coordinates by considering the spacing information. The patches from the three images are concatenated (*i.e.*, the block size is $8 \times 24 \times 24 \times 3$) and then pass through a series of three convolutional layers with 64 filters in cascade (Level 1). The output feature maps are reduced to the half of original size (*i.e.*, $4 \times 12 \times 12$) by average pooling and then pass through four convolutional layers with 128 filters in cascade (Level 2). In the same manner, the feature maps are reduced more to the size of $4 \times 6 \times 6$ and pass through five convolutional layers with 256 filters (Level 3).

The decoding stage starts from the feature maps generated at Level 3. Like the U-Net [1], we apply a deconvolution layer with a kernel $1 \times 2 \times 2$ and stride $1 \times 2 \times 2$ to generate the features maps which are compatible with the feature maps from the encoder at Level 2. The features maps from the encoder and the decoder are concatenated followed by two convolutional layers and then upsampled by a deconvolution layer. Above process is repeated to get the features maps at Level 1. For those feature maps, the last convolution operation is applied to reduce the number of feature maps to eleven which is the number of labels available in the dataset, followed by a softmax activation function. For every voxel, the index with the highest score is used as a final class prediction.

Unlike the U-Net, we add transition layers [2] between the encoder and the decoder. Specifically, the feature maps generated in the last layer of encoder at Level 1 pass through a convolutional layer with 16 filters and then are connected to the decoder at Level 1. Similarly, the feature maps on the encoder at Level 2 pass through a convolutional layer with 32 filters and then are connected to the decoder at Level 2.

All convolutional layers have $3 \times 3 \times 3$ kernels and ReLU activation and a batch normalization layer [3] are used after each convolutional layer.

2 Implementation Details

To address the intensity variations between subjects, we normalize the intensities in each patch so that the mean of intensities equals to zero and the standard deviation equals to one for each modality.

The model is trained with a weighted form of the log loss function to compensate for the size of object for every single class. In this manner, the small objects like white matter hyperintensities and the large objects like the white matter contained in the 3D patch produce the similar size of gradients. The model is trained with Adam optimizer, learning rate 1e-4 with decreases of 10 percent every $40k$ steps with a total of $500k$ training steps.

For inference, we sample the 3D patches with a stride of $4 \times 12 \times 12$ from the input set of FLAIR, T1 and IR images. Since the stride is half of the 3D patch size, the label on every voxel is predicted 8 times. The final prediction is obtained by averaging all those 8 prediction scores, before applying argmax to determine the final class of each voxel. Among 7 training images, 5 images were used for training and 2 for validation during development, but all 7 images were used to train the model for submission.

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation (NRF) of Korea funded by the Ministry of Education (2018R1D1A1B07044473)

References

1. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Proceedings of Medical Image Computing and Computer Assisted Intervention. (2015)
2. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. arXiv:1802.02611 (2018)
3. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37. ICML'15, JMLR.org (2015) 448–456