KiTS19 Submission

Satoshi Kondo[1]

[1] satoshi.kondo.jp@gmail.com

Abstract. This report describes our method submitted to 2019 Kidney Tumor Segmentation (KiTS19) Challenge. Our method employs two step approach. In the first step, an input volume is divided into two-dimensional images in three orthogonal planes and the two-dimensional images are fed into encoder-decoder networks to segment kidney and tumor regions. In the second step, the segmentation results for three orthogonal planes are fed into convolutional neural networks to obtain final segmentation results. Although our method is based on two-dimensional segmentation, but three-dimensional like processing can be performed by combining segmentation results in three orthogonal planes.

1 Our method

1.1 Overview

Our method employs two step approach. In the first step, an input volume is divided into two-dimensional images in three orthogonal planes, i.e. axial (X-Y), sagittal (Y-Z) and coronal (Z-X) planes. The two-dimensional images are fed into encoder-decoder networks to segment kidney and tumor regions. In the second step, the segmentation results for three orthogonal planes are fed into convolutional neural networks to obtain final segmentation results. We train the models in the first step and the second step separately. In the followings, preprocessing, the first step of the segmentation and the second step of the segmentation will be explained.

1.2 Preprocessing

Voxel values in an input volume are clipped between a minimum HU value and a maximum HU value, then normalized to [0, 255]. The minimum HU value is -125 and the maximum HU value is 225 in our experiment. The normalized volume is resampled to have isotropic voxels. The resampling factor is decided based on the length (in pixels) of the axial plane that the longer axis in the axial plane becomes a predetermined length. The predetermined length is 320 pixels in our experiment.

1.3 First step

In the first step, an input volume is divided into two-dimensional images in three orthogonal planes, i.e. axial (X-Y), sagittal (Y-Z) and coronal (Z-X) planes. The two-

dimensional images are fed into encoder-decoder networks to segment kidney and tumor regions from background. Figure 1 shows the whole structure of our network in the first step. We use U-Net [1] and LinkNet [2] type deep neural networks with different encoders from the original U-Net and LinkNet. U-Net and LinkNet both have an encoder-decoder structure and intermediate feature maps in the encoder are concatenated or summed to intermediate feature maps in the decoder, respectively. Our encoder is based on 101-layer ResNeXt [3] with Squeeze-and-Excitation blocks [4]. We also employ multi-task learning framework. The additional task is classification of the input image. The classes of the images are defined as followings. Class 0 is assigned to images without kidney and tumor regions, class 1 is assigned to images with kidney regions and without tumor regions, and class 2 is assigned to images with tumor regions. We add two fully connected layers on top of the last residual block of the encoder ('Residual + SE Block #4' in Fig. 1) and obtain the classification results. We also obtain the lesion area (segmentation mask) as output of the decoder. The network is trained on the tasks of classification and segmentation simultaneously.

In the training phase, images in sagittal and coronal planes are selected as the center position of kidney in Z direction comes to the center line in the image by referring to the ground truth of the segmentation. In the prediction phase, the center position is decided by referring to the prediction results of axial planes.

The training procedure is as follows. We use an encoder pre-trained on the ImageNet dataset [5]. We use an Adam optimizer [6] with an initial learning rate of 0.01. The mini-batch size is 64 and we run 100 epochs. The loss function is summation of the classification loss and the segmentation loss. The classification loss is softmax cross entropy, and the segmentation loss is the summation of pixel-wise softmax cross entropy loss and dice loss [7]. Data augmentation is applied on the fly during the training. We augment using translation, rotation, resizing, flipping, dropout and contrast adaptations. We select a model which produces the lowest loss value for validation data in the training data. In our experiment, 200 cases (volumes) are used for training and 10 cases (volumes) are used for validation. Those cases are provided by KITS19 organizers and we do not use external datasets.

In the prediction phase, we have 18 channels for each slice in a volume as a result since we have two different networks, i.e. U-Net type network and LinkNet type network, three channel segmentation result for each image and three different planes.

1.4 Second step

In the second step, the input data is the segmentation results obtained in the first step and have 18 channels for each slice. We use three-layer convolutional neural networks without down sampling and up sampling. The first and second layer has 3x3 convolution – ReLu - batch normalization structure and the number of output channels is 36. The third layer has 1x1 convolution – softmax structure and the number of output channels is 3.

The training procedure is as follows. We use an Adam optimizer [6] with an initial learning rate of 0.01. The mini-batch size is 256 and we run 50 epochs. The loss function is the summation of pixel-wise softmax cross entropy loss and dice loss [7]. Data

augmentation is not applied. We select a model which produces the lowest loss value for validation data in the training data. In our experiment, 200 cases are used for training and 10 cases are used for validation which are the same cases in the training phase.

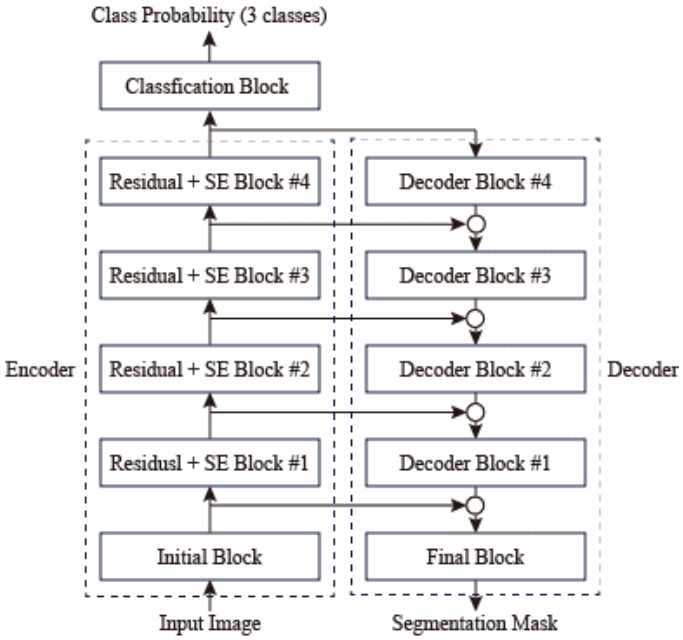

Fig. 1. Network architecture of the proposed method. 'SE' stands for 'Squeeze and Excitation'. White circles in the decoder are concatenation or summation for U-Net type and LinkNet type, respectively.

2 Experimental results

Our proposed method was evaluated with test dataset provided by the KiTS19 organizers. The results were evaluated with Dice coefficients for kidney area (including tumor area), tumor area and an average value of those coefficients. The Dice coefficients for kidney area and tumor area were 0.9324 and 0.5796, respectively. The average Dice coefficient was 0.7560.

References

1. Ronneberger, O., et al.: U-net: Convolutional networks for biomedical image segmentation. In: 18th Int. Conf. Medical Image Computing and Computer-Assisted Intervention (MICCAI), 234-241 (2015).
2. Chaurasia, A., et al.: LinkNet: Exploiting encoder representations for efficient semantic segmentation. In: IEEE Visual Communications and Image Processing (VCIP), 1-4 (2017).

3. Xie, S., et al.: Aggregated Residual Transformations for Deep Neural Networks. In: IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 1492-1500 (2017)
4. Hu, J., et al.: Squeeze-and-Excitation Networks. In: IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 7132-7141 (2018).
5. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. International Journal of Computer Vision (IJCV). 115(3), 211-252 (2015).
6. Kingma, D. P., et al.: Adam: A method for stochastic optimization. In: Int. Conf. Learning Representations (ICLR) (2015).
7. Milletari, F., et al.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 4th Int. Conf. 3D Vision (3DV). 565-571 (2016).