Multi-Encoder U-Net for Automatic Kidney Tumor Segmentation

Xueying Chen[1] and Chao Xu[1]

[1] Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, China

Abstract. Kidney tumor segmentation is a difficult yet critical task for medical image analysis. In recent years, deep learning based methods have achieved many excellent performances in the field of medical image segmentation. In this paper, we propose a Multi-Encoder U-Net segmentation method to tackle the challenging problem of kidney tumor segmentation from CT images. Our Multi-Encoder U-Net method uses three different depth networks as encoders for kidney tumor segmentation: VGG16, ResNet34, ResNet50, a feature fusion network- FED-Net is also used simultaneously, finally fusing the four results. We tested our method on the dataset of MICCAI 2019 Kidney Tumor Segmentation Challenge(KiTS).

Keywords: Kidney tumor segmentation, deep learning, Multi-Encoder

1 Introduction

Kidney tumors are one of the common causes of urinary male reproductive system tumors, and the incidence rate is high. About 95% of kidney tumors are malignant, benign is rare. Manual kidney tumor segmentation is time-consuming, difficult to reproduce, and easily influenced by personal subjective experience. Therefore, fully automatic method for kidney tumor segmentation needs to be developed. However, automatic kidney tumor segmentation is a very challenging task due to the following reasons. First, CT images often have low contrast for soft tissues and the imaging noise makes it difficult to segment the tumors. Second, the position and the shape of the tumor vary significantly. Last but not the least, the tumor boundaries are unclear and the sizes of most tumor are relatively small.

In order to solve the problems encountered in automatic segmentation, Recently, methods based on deep learning have achieved many results in the field of CT images. In this paper, we propose a Multi-Encoder U-Net segmentation method to tackle the challenging problem of kidney tumor segmentation from CT images. Our Multi-Encoder U-Net method uses three different depth networks as encoders for kidney tumor segmentation: VGG16[1], ResNet34, ResNet50[2], a feature fusion network-FED-Net[3] is also used simultaneously. Using different depths of the network as the encoder part of the U-Net architecture, which can extract multi-scale information of CT slices, it can helps to detect different sizes of targets. In addition, we also use fea-

ture fusion networks- FED-Net[3], which can effectively embed more semantic information into low-level features and improve the current feature fusion mode based on the U-Net architecture network. Finally, we use the majority majority vote ensemble to ensemble the four segmentation results, which can reduce error rate and work better to ensemble low-correlated model predictions.

2 Dataset and Preprocessing

We used the publicly available dataset of MICCAI 2019 KiTS19 Challenge as the training of our proposed method. The KiTS19 datasets contains arterial phase abdominal CT scans of 300 unique kidney cancer patients who underwent partial or radical nephrectomy. There are 210 and 90 CT scans for training and testing, respectively. All patients who underwent partial or radical nephrectomy for one or more kidney tumors at the University of Minnesota Medical Center between 2010 and 2018 were candidates for inclusion in this database. Cases that could not be found a preoperative arterial phase abdominal CT were excluded. From the remaining cases, 300 were selected at random. Each slice in all CT scans has a fixed size of 512x512 pixels The image resolutions are also different from scan to scan. In order to remove irrelevant information about other organs and tissues in the CT scans for liver lesion segmentation, we cut the the image intensity values of all CT scans to the range of [-200, 250] HU. After all CT scan HU values were truncated, we normalized all slice intensities into the range [0,1] with min-max normalization.

3 Method

In the encoder-decoder framework, the encoder gradually reduces the spatial dimension of the feature maps and the decoder gradually recovers the object details and spatial dimensions. Skip connections are usually used between the encoder and decoder, which help the decoder to better retain the details of the target, especially for situations with small sample sizes. The widely adopted U-Net architecture[4] is a typical encoder-decoder structure. Due to the structure of the organ itself is fixed and the semantic information is not particularly rich, high-level semantic information and low-level features are very important (U-Net's skip connection and U-shaped structure come in handy), therefore, U-Net network[4] outperforms many other networks in medical image segmentation. Inspired by the U-Net architecture[4], we uses three different depth networks as encoders for kidney tumor segmentation: VGG16[1], ResNet34, ResNet50[2], a feature fusion network - FED-Net[3] is also used simultaneously.

3.1 Multi-Encoder U-Net

The encoder part use three pre-trained model on the ImageNet[5] - VGG16[1], ResNet34, ResNet50 [2], separately. U-Net with VGG16[1] Encoder is similar to the

original U-Net, U-Net with an encoder based on a ResNet-type architecture refered to LinkNet[6]. In this work, we use pre-trained ResNet34, ResNet50[2]. The decoder of the network consists of several decoder blocks that are connected with the corresponding encoder block. Each decoder block includes 1 × 1 convolution operation that reduces the number of filters by 4, followed by batch normalization and transposed convolution to upsample the feature map. An overview of the network is given in the Fig. 1.

The FED-Net[3] is feature fusion encoder-decoder Network based 2D model. To effectively embed more semantic information into low-level features and improve the current feature fusion mode based on the U-Net architecture network. FED-Net[3] use a novel feature fusion method based on the attention mechanism. By assigning different weights to different features according to the contribution of the feature to the final segmentation result, more useful features can be extracted. In this work, we only used the feature fusion module of FED-Net[3], the other two modules are not used. An overview of the FED-Net[3] is given in the Fig. 2.

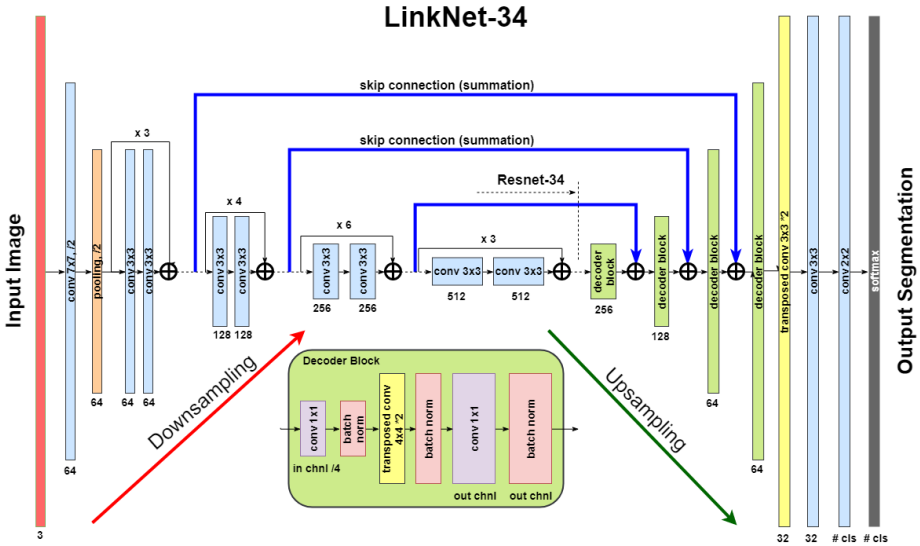

Fig. 1: The overall architecture of the LinkNet

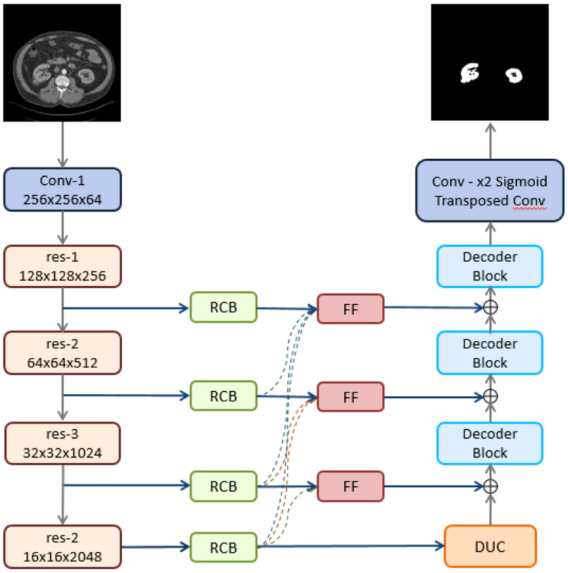

Fig. 2. The overall architecture of the FED-Net.

3.2 Hierarchical Segmentation

For kidney tumor segmentation, we take a two-step process. The first step is to segment the kidney and the second step is to segment the kidney tumor. In the first stage of the kidney segmentation, we used the slices containing only the kidney and additional 5 slices containing no kidney before and after for training and validation. In the test phase, all the slice were used. In the second stage for tumor segmentation, in order to reduce the number of negative samples and the amount of computation, we sent a slice containing only the kidney into the network. In the test phase, only the kidney slice obtained from the first step segmentation were used. To alleviating the imbalance between the positive and negative samples of the dataset, we implemented a probability extraction strategy to extract positive and negative samples from the dataset during training. The positive samples were extracted with a probability of 0.9, and the negative samples were extracted with a probability of 0.1. To make full use of the third dimension information of the CT image and without adding extra calculations, we sent the adjacent three slices as the three channels of the intermediate slice into the network.

4 Experiments

4.1 Implementation Details

Our method was implemented using the pytorch package. The experiment used a single NVIDIA GTX 1070 GPU with 8 GB memory. We used stochastic gradient

descent method for optimization with momentum of 0.9 and weight decay of 0.0001. We used the CT scans from 0 to 159 , 160 to 199 and 200 to 209 of the KiTS19 dataset as our training set, validation set and testing set respectively. Data augmentation is essential to teach the network the desired invariance and robustness properties, when only few training samples are available. In order to boost the network performance, in addition to conventional data augmentation (e.g.,image flipping, rotating, shifting, brightness adjustment and zooming) was also applied to account non-rigid deformation of the imaged organs.

In the post-processing stage: (1) The threshold for the output of kidney segmentation is 0.5, and 3D connect-component labeling was used, only the two largest 3D connected areas in the middle are reserved. (2) The threshold for the output of tumor segmentation is 0.3. On the basis of (1), a bounding box was taken for the result of kidney segmentation, and then the kidney segmentation result with bounding box was merged with the tumor segmentation result to obtain the final tumor segmentation result.

References

1. K. Simonyan and A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, arXiv:1409.1556, 2014.
2. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
3. Chen X, Zhang R, Yan P. Feature Fusion Encoder Decoder Network For Automatic Liver Lesion Segmentation[J]. arXiv preprint arXiv:1903.11834, 2019.
4. Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in International Conference on Medical image computing and computer-assisted intervention. Springer, 2015, pp. 234–241.
5. [10] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z.Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg and Li Fei-Fei,ImageNet Large Scale Visual Recognition Challenge, arXiv:1409.0575, 2014.
6. Chaurasia A, Culurciello E. Linknet: Exploiting encoder representations for efficient semantic segmentation[C]//2017 IEEE Visual Communications and Image Processing (VCIP). IEEE, 2017: 1-4.