Segmentation of kidney and kidney tumor by cascaded fusion FCNs with soft-boundary regression

Jian Zhang, Kelei He, Tiexin Qin, Jianrong Chen, Lihe Yang

Nanjing University, Nanjing, P. R. China

Abstract. To produce reliable kidney and kidney tumor semantic segmentation, we proposed a two-stage method to automatically segment kidney and tumor. Specifically, in the first stage, to crop input into a small region, we train a small network to locate kidney and tumor with down-sampled image. In second stage, we train three types of networks to segment kidney, tumor, kidney and tumor respectively. Then we combine these networks together with ensemble method to produce reliable kidney and tumor segmentation. Our method can achieve an overall approximate score of 85.1% in DSC in Kits19 Challenge, with 96.9% for kidney and 73.3% for kidney tumor.

Keywords: kidney segmentation · ensemble

1 Introduction

Delineation of kidney and kidney tumor in CT images is an important prerequisite procedure in kidney cancer treatment. Typically, manual labeling is adopted in common clinic situations and often takes experienced clinicians considerable time. Therefore, automatic kidney and kidney tumor detection and segmentation via machine learning technique is in high demand. However, the task is challenging due to 1) the unclear organ boundaries, 2) the variance of organ appearance, and 3) the relatively small size of tumor in many cases. In [], several machine learning based methods are proposed to tackle this problem. Since these methods use hand-crafted features, the problems are not perfectly solved. As deep learning-based methods shown impressive results in computer vision tasks, researchers also bring this technique into the rich field of medical image analysis [], including kidney and kidney tumor segmentation []. Despite the greatly improvement of the segmentation performance, the existing methods are not specifically designed for the kidney and kidney tumor segmentation task, thus limited their ability of distinguish such structures. In this paper, we proposed a two-stage deep framework to automatically segment these two structures (i.e., kidney and kidney tumor) from the raw CT image. Specifically, the first stage is designed for fast localization of the kidney and kidney tumor region. We use down-sampled images and do binary classification (i.e., consider both kidney and kidney tumor as foreground) to quickly and robustly generate the target region. The predicted

target region can cover both kidney and kidney tumor in original image is then cropped for fine segmentation in the second stage. We leverage different models to segment the two structures as they are of unequal task difficulty. For kidney, we use one V-Net trained on patches cropped in the kidney-tumor region. For kidney tumor, we construct a fusion network which incorporates predictions from three networks. Specifically, a 2.5D and a 3D multi-task network is constructed with the aim to segment the tumor and regress the tumor boundaries. As the tumors are hard to be distinguished, we further construct another 3D network with the aim to segment both the tumor and kidney to constrain the final results. Our method can achieve an overall score of 85.1% in DSC in Kits19 Challenge [], with 96.9% for kidney and 73.3% for kidney tumor.

2 Method

2.1 Stage 1 : kidney and kidney tumor localization

In the first stage, we localize kidney and kidney tumor in the down-sampled images with one V-Net-based segmentation network. The reasons of design the first stage are: 1) The image size of the cases are very large, and cannot be completely fed into the network at once. 2) The positive (i.e., denote the kidney or tumor) and the negative (i.e., denote the background) pixels are extremely imbalanced in the raw CT images. 3) The efficiency can be improved by using down-sampled images for quickly localizing the organ region. Thus following [Han+19], we first train a v-net with patches taken from down-sampled images to segment kidney and tumor simultaneously. Note that we consider this task as a binary segmentation task, in which kidney and kidney tumor are both regarded as foreground. This benefits the network in generating robust organ regions, as the network avoids to distinguish the border between kidney and kidney tumor which is fuzzy in down-sampled images.

To help accurately localize kidney and kidney tumor, we construct a multi-task network, in which we add a branch to regress the landmark of tumor and kidney besides segmentation of these two structures. Specifically , the branch output five-channel heatmaps for predicting five landmarks. We place the landmarks of the left, top, right, bottom and center point in each slice for kidney. As these landmarks are all on the boundaries of the kidney and tumor, they can provide strong location guidance for the network.

The kidney-tumor region is determined with bounding boxes from the segmentation network. Since the we use down-sampled images in this stage, we scale up the calculated bounding boxes to the original scale. Then the region is cropped from original images with the bounding boxes.

However, this simple cropping method may fail in some specific cases. For example, patients in some cases only have one kidney, and the network still predicts bounding boxes which contain two kidneys, and vise versa. To solve the problem, we propose a mixed cropping method, which leverages the landmark information and the prediction of kidney segmentation network in stage 2. To

solve the problem of less prediction of the bounding boxes, we first compare the width in Transverse view of two bounding boxes generated from the rough segmentation and the landmark, which are predicted by the localization network. If the width of the two predicted bounding boxes are about the same, we use the bounding box generated from the segmentation; otherwise we use the bounding box from the landmark. This comparison can prevent the network from miss prediction of the kidney, as the bounding box of landmark can always contain two kidneys. Then, we solve the problem of over prediction of the bounding boxes by using the kidney prediction network trained in stage 2. This network can generate refined kidney segmentation. We then empirically remove the small area in the predicted segmentation maps with a threshold of 1500. The kidney bounding box is generated from this refined kidney segmentation. We compare it with the previously obtained bounding box. We use the kidney bounding box in the case of its size is smaller than half of the previously obtained one.

2.2 Stage 2 : Multi-task kidney and kidney tumor segmentation network

In the second stage, we should train networks that can segment kidney and tumor inside cropped image. Since these networks does not need to learn the information outside the image, we ought to crop the image with ground truth bounding box generated from ground truth segmentation. But if the bounding box generated from stage 1 is not very close to ground truth, the area outside the ground truth will confuse the network. Thus we expand the ground truth bounding box to some extend so that it can handle more complicated situation.

Also, like the first stage, to help segment better, we add a branch to predict the boundary map of objects(kidney, tumor, or kidney and tumor). For different task, we produce different boundary map to predict. So that more information can be got.

We cropped image with the bounding box of tumor and kidney and trained three types of v-net to predict tumor (denoted as T), kidney (denoted as K) , kidney and tumor(denoted as K_T) respectively. And found that the performance of predicting both tumor and kidney is 95%+, predicting kidney is also 90%+. But the performance of predicting tumor is very bad, under 65%. One of this reason might be the imbalance of positive and negative. Thus we further only cropped the image with the bounding box of tumor. This time the performance is 75%+.

Other than 3D V-Net, we also trained 3D U-Net and and 2.5D U-Net for tumor prediction. The 2.5D U-Net is composed of three 2D U-Net trained with slices along different axis. We train these two network for two reasons. First, the performance of 3D V-Net is not good, we want to try other models to seek better networks. Second, we want to use ensemble method to predict final labels, so we need different models with different views to complement each other.

2.3 Stage 2 : model ensemble

We totally have 12 trained models. One 3D V-Net for bounding box prediction, one 3D V-Net for K_T prediction, two 3D V-Net for K prediction, one 3D V-Net , one 3D U-Net and and six 2D U-Net(3 of which are trained with dice loss, others are trained with BCE loss) for 2.5D tumor prediction. Then how to ensemble these large number of models is a big question.

Since the performance of T is very bad, and the performance of K_T an K is very good. Thus we came up with a label assignment strategy to ensemble models. First, we use K_T to predict both kidney and tumor, then we use K and T to predict their own label separately. Finally we assign labels from K and T to the labels from K_T. Specifically, for a pixel to be labeled, if K_T predicts 1 and K predicts 1, then its final label is 1(kidney). If K_T predicts 1 and T predicts 1, then its final label is 2(tumor). If both K and T predict 1, then we will label it as 2 because tumor is very hard to prediction, any wrong labeled pixel will damage performance.

In addition, when ensemble same networks like 3 tumor prediction networks, for same network but different training data, we average their output as one ensemble model, then for different ensemble networks we we directly use the strategy above for each model to get final output. We do not need to worry too much about the overwhelming tumor label because the constrain of K_T prediction can eliminate most false positive labeling.

Finally, we pad the label of cropped image to its original size with zero to get final prediction.

3 Experiments and results

3.1 Dataset

The data is from Kits19 training and testing data sets. The training data and testing data include 210 and 90 patients, respectively. We use its interpolated version. All data in this version are resampled to $3 \times 0.78 \times 0.78$. The depth varies from 145 to 755 and the spatial size varies from 224 to 533.

3.2 Data preparation

We first clip image to [-200, 300], then normalize it to [0, 1]. For the 3D V-Net and U-Net, we first crop image and then take random or stride patches from the cropped image. In first stage, the landmark is generated on downsampled image. For each slice along z axis, we first get 5 coordinates of left, top, right, bottom and center of a kidney using segmentation for each kidney. Then put 1 on an empty slice with corresponding coordinate. After that, we put a gaussian filter on the slice with sigma of 3 to ease regression and divide the slice with maximum value. Finally we put all landmark slices together to get a landmark ground truth for a image. In second stage, the boundary is generated from resampled segmentation. The procedure is the same as landmark generation except for that

we need to get a boundary using segmentation and the sigma of the gaussian filter is 3 for kidney and 2 for tumor. Then we can take patches from landmark or boundary to generate training data.

Specifically, when crop image, for bounding box prediction, we first downsample image with factor of 4, then crop 100 random patches with size of $16 \times 64 \times 64$ in the whole downsampled image. For K_T and K prediction, we first crop image around kidney and tumor area with expand size of $16 \times 128 \times 128$, then crop 200 stride patches in it. For T prediction, we first crop image around tumor area with expand size of $8 \times 16 \times 16$, then take 100 random patches with size of $32 \times 64 \times 64$. For 2.5D tumor prediction, we only use slices containing tumor as training data to ease class imbalance, and random crop image with size of 256×256, for images that smaller than this size, we pad it with zero.

3.3 Model training

The network for predicting bounding box is 5 level V-Net with an additional branch starting from bottom level to the highest level. Its landmark output channel is 5. The K_T network and one of K network is the same structure except for the channel is 1 for boundary regression. Another K network has a branch starting from second last level instead of bottom level to share more common information. Two of the T networks have the same structure and another is a 4 level 3D U-Net.

We trained all network with dice loss except three 2D U-Net which adopted BCE loss. We use weighted MSE(WMSE) loss for boundary and landmark regression :

$$WMSE(l,t) = \left(\frac{\sum_{\{p|t_p>0\}}(l_p - t_p)^2}{\sum_{\{p|t_p>0\}} t_p} + \frac{\sum_{\{p|t_p<=0\}}(l_p - t_p)^2}{\sum_{\{p|t_p<=0\}} t_p}\right)/2$$

where l is prediction, t is ground truth, p is an index in prediction. We adopt poly scheduler $lr * (1 - \frac{iter}{total\_iter})^{0.9}$, adam optimizer with weight decay of 5e-4. The learning rate for K_T and K networks is 1e-4, for T network is 1e-3.

3.4 Post process

For tumor prediction, it is easy to produce small noise. We pick connected components with size > 64 in each 2D slice to remove these slice.

3.5 Result

We evaluate our ensemble model online using 90 testing data. Table 1 shows the dice accuracy.

	kidney and tumor	tumor	average
Dice	0.963	0.713	0.838

4 Conclusion

In conclusion, we adopt a cascaded method to first locate kidney and tumor, then segment them with ensembled models, where we propose a label assignment strategy to deal with the difficulty of segmenting tumor.

References

[Han+19] Miaofei Han et al. "Segmentation of CT Thoracic Organs by Multi-resolution VB-nets." In: *SegTHOR@ ISBI*. 2019.