Automatic Kidney and Tumor Segmentation with Hybrid Hierarchical Networks

Yading Yuan

Department of Radiation Oncology
Icahn School of Medicine at Mount Sinai
New York, NY, USA

Abstract. Automatic segmentation of kidney and its tumors is an essential but challenging step for extracting quantitative imaging biomarkers for accurate tumor detection, diagnosis, prognosis and treatment assessment. Kidney Tumor Segmentation Challenge (KiTS) provides a common platform for comparing different automatic algorithms on abdominal CT images in tasks of 1) kidney segmentation and 2) kidney tumor segmentation . We participate this challenge by developing a fully automatic framework based on deep neural networks. By observing that clinicians usually contour organs and tumors in the axial view while evaluating the contours in 3D space, we adopt a 3-step hierarchical structure with hybrid 2D and 3D models. In the first step, a simple 2D U-Net model is trained to obtain a quick but coarse segmentation of the kidney region on the entire 3D CT volume; then another 2D U-Net using residual blocks with channel-wise attention is applied to each kidney region for kidney and tumor segmentation. At last, the segmented tumor is refined by a 3D model for final tumor segmentation. Our framework was trained using the 210 challenge training cases provided by KiTS. By 5-fold evaluation, our method achieved an average Dice Similarity Coefficient (DSC) of 0.970 on kidneys and 0.756 on kidney tumors, respectively.

1 Introduction

Kidney cancer is one of the most rapidly increasing cancers in terms of incidence and mortality worldwide, raising from 208,000 diagnoses and 102,000 deaths in 2002 to more than 400,000 diagnoses and 175,000 deaths in 2018 (1; 2). Although the increasing use of CT abdominal imaging has allowed kidney tumor be detected, diagnosed, and treated in the early stage, which has contributed to the disease's increased overall survival (3), proper interpretation of CT images is normally time-consuming and prone to suffer from inter- and intra-observer variabilities. Meanwhile, the current scoring systems, such as RENAL and PADUA, can only characterizes relatively simple and easy-to-extract tumor features from CT images, which limits the predictive power of imaging studies. As the result, computerized analysis have been of great demand to assist clinicians for better interpretation of abdominal CT images for kidney cancer. Specially, automatically segmenting kidney and viable tumors from other tissue is an essential step in quantitative image analysis of abdominal CT images.

In order to accelerate the research and development of reliable methods for automatic kidney and kidney tumor segmentation, MICCAI 2019 Kidney Tumor Segmentation Challenge (KiTS) provides a common platform for comparing different automatic algorithms on abdominal CT images in tasks of 1) kidney segmentation and 2) kidney tumor segmentation . We participate this challenge by developing a fully automatic framework based on deep neural networks. By observing that clinicians usually contour organs and tumors in the axial view while evaluating the contours in 3D space, we adopt a 3-step hierarchical structure with hybrid 2D and 3D models. In the first step, a simple 2D U-Net model is trained to obtain a quick but coarse segmentation of the kidney region on the entire 3D CT volume; then another 2D U-Net using residual blocks with channel-wise attention is applied to each kidney region for kidney and tumor segmentation. At last, the segmented tumor is refined by a 3D model for final tumor segmentation.

2 Datasets and preprocessing

Only KiTS challenge datasets (4) were used for model training and testing. The KiTS datasets consist of 300 multi-phase abdominal CT images provided by University of Minnesota, in which 210 cases were used for training and the rest of 90 for testing. The datasets have significant variations in image quality, spatial resolution and field-of-view, with in-plan resolution ranging from 0.44 × 0.44 to 1.04 × 1.04 mm and slice thickness from 0.5 to 5.0 mm. Each axial slice has identical size of 512 × 512, but the number of slices in each scan varies from 29 to 1059.

As for pre-processing, we simply truncated the voxel values of all CT scans to the range of [-135, 215] HU to eliminate the irrelevant image information. This HU range is the default setting for reviewing abdominal CT images in 3D Slicer (www.slicer.org). Our 2D models are based on 2D slices and the CT volume is processed slice-by-slice, with the two most adjacent slices concatenated as additional input channels. Different resampling strategies are applied at different hierarchical levels and will be described below.

3 Methods

3.1 Basic network structure

We use U-Net (5) as the basic network structure of our framework, as shown in Fig. 1. Convolution and max-pooling are employed to aggregate contextual information of CT images in the encoding pathway, and transpose convolution is used to recover the original resolution in the decoding pathway. Each convolutional layer is followed by batch normalization and rectified linear unit (ReLU) to facilitate gradient back-propagation. Long-range skip connections, which bridge across the encoding blocks and the decoding blocks, are also created to allow high resolution features from encoding pathway be used as additional inputs to the convolutional layers in the decoding pathway.

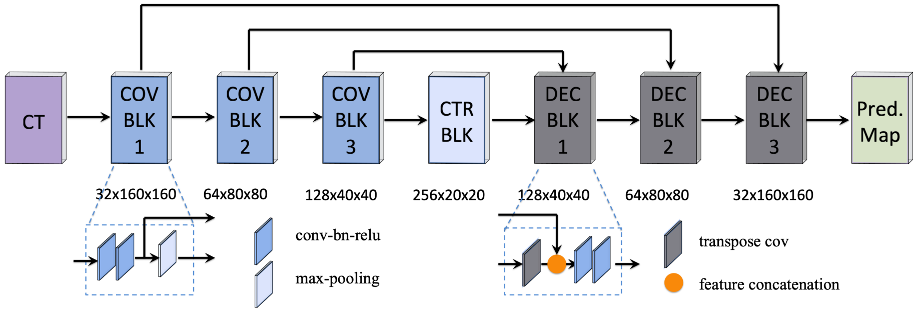

Fig. 1. Architecture of U-Net. This architecture employs convolution and max-pooling to aggregate contextual information, and uses transpose convolution and long-range skip connection for better determination of seed locations. The numbers under each block represent the dimensions of its feature output, in which the first dimension denotes the feature channel.

3.2 Localizer network

This network is used to locate the kidney regions by performing a fast but coarse kidney segmentation on the entire CT volume, thus we just use a simple U-Net for this purpose. This model includes 100 layers and 2.9 M trainable parameters and its architectural details can be found in Fig. 1. For each CT volume, the axial slice size was firstly reduced to 160 × 160 by down-sampling and then the entire image volume was resampled with slice thickness of 3 mm. We found that not all the slices in a CT volume were needed in training localizer, so only the slices with kidneys, as well as the 2 slices superior and inferior to the kidneys were included in the model training. The kidney and tumor labels were merged as a single kidney label to provide the ground truth during model training. A simple square error was used as loss function.

During testing, the new CT images were pre-processed following the same procedure as training data preparation, then the trained localizer network was applied to each slice of the entire CT volume. Once all slices were segmented, a threshold of 0.5 was applied to the output and a 3D connect-component labeling was performed. A volume size threshold of 100 ml was used and the largest one or two connected component were selected as the initial kidney regions.

3.3 Segmenter network

An accurate kidney localization enables us to perform a fine kidney and tumor segmentation with more advanced model while reducing computational time. Specifically, we first resampled the original image to 0.6 × 0.6 × 2.0 mm, then extracted a 256 × 256 region of interest (ROI) from the kidney location in every slice within 5 slices of the kidney region. Besides the original image intensity, a 3D regional histogram equalization was implemented to enhance the contrast between tumors and the surrounding kidney tissues, in which only those voxels inside the initial kidney mask were considered in constructing intensity histogram. Note that we processed each kidney independently, resulting in around 27,000 training samples for model training.

We enhanced the U-Net in the following four aspects: 1) We replaced the original convolution blocks with residual blocks (6) to allow a better information flowing, each of which includes three convolution layers; 2) We added a squeeze-excitation layer (7) at the end of each residual block in the encoding pathway to calibrate the channel-wise response; 3) We added deep supervision at each level of decoding pathway to improve the training stability; 4) We generalized the Jaccard distance loss (8), which we developed in our previous work for single object segmentation, to multiple objects, and combined it to cross entropy as the loss function of segmenter training. This model includes 253 layers with about 20 M trainable parameters.

During testing, kidney VOI was extracted based on the initial kidney mask obtained from localizer network, then the trained segmenter was applied to each slice in the VOI to yield a 3D probability map. A testing time augmentation was also applied for better segmentation.

3.4 Refiner network

In the past two steps, we employed 2D models to emphasize the contextual information integration in axial planes. While the regional information in z (superior/inferior) direction can be incorporated into the models by adding the most two adjacent planes as additional input features, this integration is not sufficient to catch a larger scale contextual information in this direction.

We extended the 2D segmenter network to a fully 3D model to further refine the tumor segmentation. We resampled the original image to $0.6 \times 0.6 \times 2.0$ mm, and extracted a $32 \times 128 \times 128$ volume of interest (VOI) from each tumor candidate that was generated from the segmenter. The VOI was centered at the centroid of tumor candidate and a sliding-window strategy with stride of 16 in z direction was used if the tumor candidate can not be fully covered by VOI. Meanwhile, the tumor candidate mask itself also served as an additional input channel to refiner model. Eventually, we obtained over 700 VOIs for model training.

Considering the number of VOIs and computational resource available for model training, we made the following modifications on the segmenter network structure: 1) We employed an anisotropic down-sampling method to perform max-pooling only in $x - y$ plane in the first two residual blocks; 2) We reduced the number of convolution layers in each residual block from three to two; 3) We removed deep supervision in the decoding pathway; 4) We replaced the batch normalization with group normalization (9) considering that only small batch size can be used for 3D operations. This model includes 146 layers with about 11 M trainable parameters.

During testing, the tumors obtained from the refiner network were added to the results from segmenter to yield a final segmentation.

3.5 Implementation

Our framework was implemented with Python using Pytorch (v.0.4) package. Training each network took 300 iterations from scratch using Adam stochastic optimization method. The initial learning rate was set as 0.003, and learning rate decay and early stopping strategies were utilized when validation loss stopped decreasing. The batch size was 16 for 2.D model training and 4 for 3D model. In order to reduce overfitting, we randomly flipped the input volume in left/right, superior/inferior, and anterior/posterior directions on the fly for data augmentation. We used 5-fold cross validation to evaluate the performance of our model on the training dataset, in which a few hyper-parameters were also experimentally determined via grid search. All the experiments were conducted on a workstation with four Nvidia GTX 1080 TI GPUs.

4 Experiments

Since the ground truth of testing dataset is not released, here we report the results from 5-fold cross validation. Table 1 shows the segmentation results in terms of Dice similarity coefficient (DSC) for localizer network, and table 2 is for segmenter network. Table 3 shows the results after refiner network. When applying the trained models on the challenge testing dataset, a bagging-type ensemble strategy was implemented to combine the outputs of five refiner networks to further improve the segmentation performance.

Table 1. Segmentation results (DSC) of localizer network in 5-fold cross validation.

	fold-0	**fold-1**	**fold-2**	**fold-3**	**fold-4**
Kidney	0.949	0.948	0.952	0.939	0.945

Table 2. Segmentation results (DSC) of segmenter network in 5-fold cross validation.

	fold-0	**fold-1**	**fold-2**	**fold-3**	**fold-4**
Kidney	0.967	0.966	0.966	0.952	0.964
Tumor	0.746	0.702	0.701	0.643	0.727
Average	0.857	0.834	0.834	0.798	0.844

Table 3. Segmentation results (DSC) of refiner network in 5-fold cross validation.

	fold-0	**fold-1**	**fold-2**	**fold-3**	**fold-4**
Kidney	0.970	0.968	0.972	0.968	0.9701
Tumor	0.795	0.737	0.775	0.732	0.742
Average	0.882	0.853	0.874	0.850	0.856

Bibliography

[1] Bray, F., et al.: Global cancer statistics 2018: Global estimates of incidence and mortality worldwide for 30 cancers in 185 countries. Cancer J Clin. 68(6), 394-424, 2018.

[2] Parkin, D. M., et al.: Global cancer statistics 2002. Cancer J Clin. 55(2), 74-108, 2005.

[3] Homma, Y., et al.: Increased incidental detection and reduced mortality in renal cancer: recent retrospective analysis at eight insitutions. International Journal of Urology 2(2), 77-80, 1995

[4] Heller, N., et al.: The KiTS19 Challenge Data: 300 kidney tumor cases with clinical context, CT semantic segmentations, and surgical outcomes. arXiv:1904.00445.

[5] Ronneberger, O., et al.: U-Net: Convolutional networks for biomedical image segmentation. in Proc. MICCAI 2015. Springer, 234-241, 2015.

[6] He, K., et al. Deep residual learning for image recognition. In proc. CVPR 2016, 770-778, 2016.

[7] Hu, J., et al. Squeeze-and-excitation networks. In proc. CVPR 2018, 7132-7141, 2018.

[8] Yuan, Y., et al. Automatic skin lesion segmentation using deep fully convolutional networks with Jaccard distance. IEEE Trans. Med. Imaging, 36(9), 1876-1886, 2017.

[9] Wu, Y., et al. Group normalization. In proc. ECCV 2018, 3-19, 2018.