Coarse to Fine Framework for Kidney Tumor Segmentation

Shuolin Liu

Department of Electrical Engineering and Automation, Anhui University, Hefei
230601, China
z18201021@stu.ahu.edu.cn

Abstract. Accurate segmentation of kidney tumor is a key step in image-guided radiation therapy. However, shapes, scales and appearance vary greatly from patient to patient, which pose a serious challenge to segment targets correctly. In this work, we proposed a coarse-to-fine framework to automatically segment kidney and tumor computed tomography (CT) images. Specifically, we adopt two resolutions and propose a improved 3D U-Net network for kidney tumor segmentation. The model in the coarse resolution can robustly localize the kidney, while the model in the fine resolution can accurately refine the boundary of kidney and tumor.

Keywords: kits19

1 Introduction

There were more than 400,000 kidney cancer diagnoses worldwide in 2018 resulting in more than 175,000 deaths , up from 208,000 diagnoses and 102,000. The morphometry of a kidney tumor revealed by contrast enhanced Computed Tomography (CT) imaging is an important factor in clinical decision making surrounding the lesions diagnosis and treatment. In this work, we proposed coarse to fine framework to automatically segment kidney and tumor computed tomography (CT) images. Specifically, we adopt two resolutions and propose a improved 3D U-Net. [1] network for kidney tumor segmentation. The model in the coarse resolution can robustly localize the kidney, while the model in the fine resolution can accurately refine the boundary of each kidney and tumor.

2 Method

2.1 Multi-resolution strategy

Many deep learning algorithms segment organs using a single resolution. As 3D medical images are often large in size, e.g. 512×512×200, passing the whole 3D image volume into networks will consume a lot of GPU memory, hence increasing the chances of segmentation failure due to lack of GPU memory. In this work, we

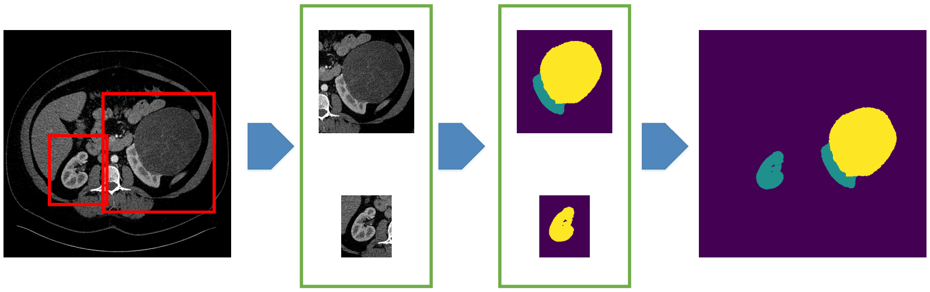

Fig. 1. An overview of our model architecture

adopt a multi-resolution strategy. In the coarse segmentation stage, we trained our model using resampled images at 1.62 × 1.62 × 3.21 mm voxel size, resulting in a median image shape of 128×248×248 voxels for the training cases. In the fine segmentation stage, we trained our model only using the left/right VOI (defined as the region of kidney dilated of 20mm) with the voxel spacing of 0.78 × 0.78 × 3.0. Each case is clipped to the range [-90, 310]. We then subtract 110 and divide by 200 to bring the intensity values in a range.

2.2 Network for Kidney Tumor Segmentation

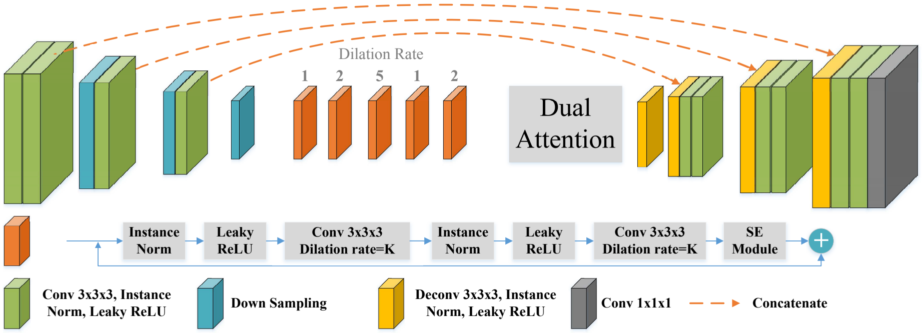
Dilation Rate
1
2
5
1
2
Dual Attention
Instance Norm
Leaky ReLU
Conv 3x3x3 Dilation rate=K
Instance Norm
Leaky ReLU
Conv 3x3x3 Dilation rate=K
SE Module
Conv 3x3x3, Instance Norm, Leaky ReLU
Down Sampling
Deconv 3x3x3, Instance Norm, Leaky ReLU
Conv 1x1x1
Concatenate

Fig. 2. An overview of our model architecture

Our model is adapted from 3D Unet [1], we employ squeeze-excitation hybrid dilated convolutions (SE-HDC) and volumetric dual attention to effectively enlarges the receptive fields and aggregate global information, which is shown in Fig. 2.

Table 1. Detials of model architecture

Module	Kernel size	Stride	Output size (Coarse / Fine)
Input	\	\	2×1×160×160×80 / 2×1×144×144×48
Conv	3×3×3	1×1×1	2×24×160×160×80 / 2×24×144×144×48
Down Sampling	3×3×3	2×2×1	2×48×80×80×80 / 2×24×72×72×48
Conv	3×3×3	1×1×1	2×48×80×80×80 / 2×24×72×72×48
Down Sampling	3×3×3	2×2×2	2×96×40×40×40 / 2×96×36×36×24
Conv	3×3×3	1×1×1	2×96×40×40×40 / 2×96×36×36×24
Down Sampling	3×3×3	2×2×2	2×192×20×20×20 / 2×192×18×18×12
SE-HDC	3×3×3	1×1×1	2×192×20×20×20 / 2×192×18×18×12
Dual Attention	3×3×3	1×1×1	2×192×20×20×20 / 2×192×18×18×12
Up Sampling	2×2×2	2×2×2	2×96×40×40×40 / 2×96×36×36×24
Concat	\	\	2×192×40×40×40 / 2×192×36×36×24
Double Conv	3×3×3	1×1×1	2×96×40×40×40 / 2×96×36×36×24
Up Sampling	2×2×2	2×2×2	2×48×80×80×80 / 2×24×72×72×48
Concat	\	\	2×96×80×80×80 / 2×96×72×72×48
Double Conv	3×3×3	1×1×1	2×48×80×80×80 / 2×24×72×72×48
Up Sampling	2×2×1	2×2×1	2×24×160×160×80 / 2×24×144×144×48
Concat	\	\	2×48×160×160×80 / 2×48×144×144×48
Double Conv	3×3×3	1×1×1	2×24×160×160×80 / 2×24×144×144×48
Final Conv	1×1×1	1×1×1	2×1×160×160×80 / 2×1×144×144×48

Squeeze-Excitation Hybrid Dilated Convolutions In U-like convolutional neural network, a stack of down-sampling operations are used to reduce the resolution of feature maps and achieve larger receptive field, which poses serious challenges to preserve the details. To address this problem, we employed squeeze-excitation hybrid dilated convolutions to enlarge the receptive fields and capture dense features. We followed the idea of Yang et al to avoid gridding issue [2] when we designed the block with hybrid dilated convolutions (HDC). We draw our inspiration from the recently proposed squeeze-excitation (SE) modules for channel recalibration for image classification. The dilation rates of 3×3 squeeze-excitation convolutions are repeated by a sequence of 1, 2, 5, 1, 2, their corresponding receptive field could varies to 3×3×3, 5×5×5, 11×11×11, 3×3×3, 5×5×5. This mechanism allows us to effectively enlarge the receptive field without down-sampling operations. Besides, dilated convolutions have the same number of parameters as the original 3×3×3 convolutions. Let denote the 3×3×3 convolution kernel that related to layer $\mathcal{L}^\ell$ by $\mathcal{K}^\ell$, then discrete convolutions could describes as follow:

$$(\mathcal{L}^\ell *_d \mathcal{K}^\ell)(\boldsymbol{p}) = \sum_{a+db=\boldsymbol{p}} \mathcal{L}^\ell(a)\mathcal{K}^\ell(b) \tag{1}$$

where $\boldsymbol{p}$ is the the domain of feature maps in $\mathcal{L}^\ell$, $*_d$ is the discrete convolution operator with dilation rate of d.

Dual Attention We extand dual attention module, a powerful tools in recent semantic segmentation, to volumetric segmentation task. A position attention and channel attention module is proposed to learn the spatial interdependencies of features and a channel attention module is designed to model channel interdependencies. It significantly improves the segmentation results by modeling rich contextual dependencies over local features. Detials could be found in [3]

2.3 Training Procedure

The sum of the cross-entropy loss and the dice loss are used as loss function. Data Augmentation includes elastic deformations, random scaling and random rotations as well as gamma augmentation. Adam was used as optimizer for stochastic gradient descent with the batch size of 2. The initial learning rate is 0.0003 and l_2 weight decay is 1e-5. Whenever the exponential moving average of the training loss does not improve within the last 30 epochs the learning rate is dropped by a factor of 0.2. Training is stopped when the learning rate drops below 1e-6 or 700 epochs are exceeded.

2.4 Inference Procedure

Cases are predicted using a sliding window approach with half the patch size overlap between predictions. Then, we pick the top 2 largest connected component as the final result.

3 Result

We achieve 0.9742 dice score for kidney and 0.8231 for tumor with a single model.

Table 2. Mean dice of the proposed pipeline on KiTS19 test dataset

Network architecture	Kidney Dice	Tumor Dice	Mean Dice
Ours	0.9742	0.8231	0.8987

4 Discussion

In conclusion, we proposed a coarse-to-fine framework to automatically segment kidney and tumor computed tomography (CT) images. The model in the coarse resolution can robustly localize the kidney, while the model in the fine resolution can accurately refine the boundary of each kidney and tumor. We also tried to use dense block [4] with hybrid dilated convolutions to catch dense features, but it did not work well (0.972 for kidney, 0.805 for tumor). Due to limitations in computing resources, i submit the result using a single model. I am confident that the ensemble of proposed model would achieve higher ranking in this challenge.

References

1. Cicek, O., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. pp. 424C432. Springer (2016)
2. Wang, P., Chen, P., Yuan, Y., Liu, D., Huang, Z., Hou, X. Cottrell, G.:Understanding Convolution for Semantic Segmentation. In: IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1451–1460 (2018)
3. Fu, J. , Liu, J. , Tian, H. , Li, Y. , Bao, Y. , Fang, Z.: Dual attention network for scene segmentation(2018).
4. Huang, G., Liu, Z., Maaten, Laurens, V. D. M., Weinberger, K.Q.:Densely Connected Convolutional Networks.In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261–2269 (2017)