Dense Pyramid Context Encoder-Decoder Network for Kidney Lesion Segmentation

Zhen Yu, Youyi Song, Jing Qin

School of Nursing, HK PolyU
harry.qin@polyu.edu.hk

Abstract. In this manuscript, an automated solution is presented for the kidney lesion segmentation. The proposed method consists of two-stage learning procedures which generating prediction masks for kidney and lesion respectively. Since we adopt 2D axial images from CT scans as evaluation data, it is critical to extract sufficient contextual information for capturing the objects varied significantly in appearance within different slices. Hence, we redesign an encoder-decoder network for more effective feature representations learning. We evaluate our method on 2019 Kidney Tumor Segmentation Challenge. There are total 210 labeled CT scans released as training and validation data. The source code can be found at: https://github.com/Zakiyi/kits_2019_segmentation_challenge.

Keywords: Kidney cancer, CT image, lesion segmentation, deep learning.

1 Introduction

Kidney cancer, characterized by malignant tumor arising from the renal parenchyma and renal pelvis, is a common form of cancer affecting adults [1]. This disease, however, is highly curable when treated in the early stage. Clinically, imaging test such as CT scans is an important method for identifying kidney tumor or abnormality. Automated delineation of kidney and lesion within images can be of immense help in pre-surgical planning for the treatment, because useful information like tumor size, shape, etc., can be obtained. Indeed, over the past few decades, a considerable amount of studies have been devoted to develop algorithms toward intelligent kidney lesion segmentation. Nevertheless, artifacts, large inhomogeneity of the kidney (cortex and medulla), and similar intensities of adjacent organs, pose huge challenges in the development of accurate image analysis system.

Recently, significant improvements in medical image segmentation have been obtained by using deep convolutional neural network (CNN). However, only few studies exploited the deep learning based method for kidney CT image segmentation till now. The crucial consideration of applying CNN in dense predication lies in satisfying simultaneously the demands of multi-scale reasoning and full-resolution output [2]. In this regard, the most successful architecture in current field of medical image segmentation is U-Net [3], which constructed with an encoder path and a decoder path. Hierarchical features are first learned by encoder path, and then decoder path gradually

recovers detail information by fusing counterpart features from encoder path via skip connections.

In this article, we present a dense pyramid context encoder decoder network for kidney lesion segmentation based on original U-Net. Although most existing studies using pre-trained model from natural image classification task as the backbone of feature encoding path and superior performance can be achieved by fine-tuning, these

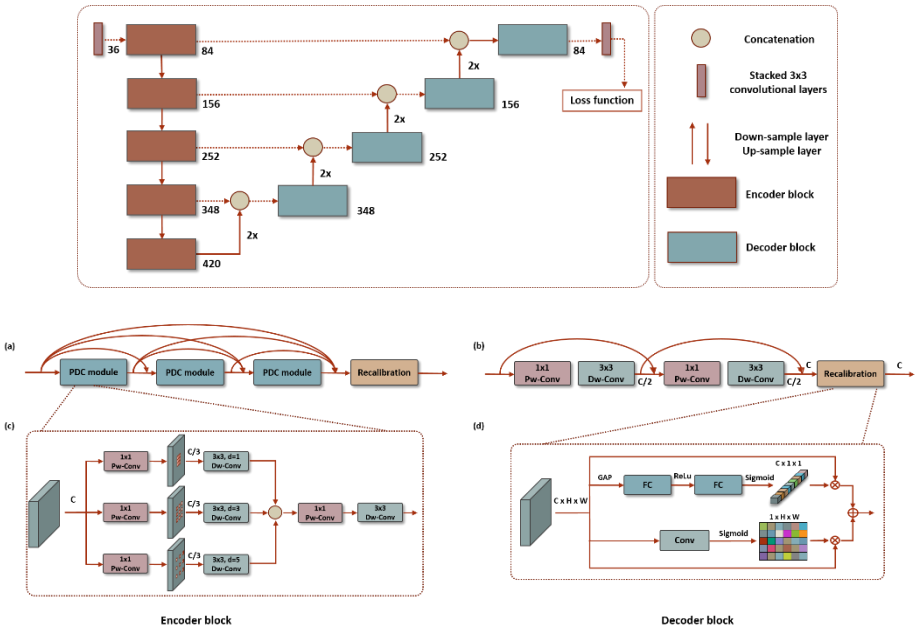

Fig. 1. Proposed network architecture for the kidney lesion segmentation, the entire model is similar to U-net, while we re-designed the basic components of encoder part and decoder part. Fig 1 (a) represents the encoder block, which consists of several pyramid dilated convolutional models (PDC model) and a recalibration model; Fig 2 (b) illustrates the architecture of the decoder block, each of them incudes two separable convolution layers and a recalibration model; The construction of PDC model and recalibration model was presented in (c) and (d) respectively.

methods have a restricted network designing space. In contract, the proposed model incorporates a series of popular designing elements in computer vision includes dense connection, separable convolution, pyramid dilated convolution and feature recalibration based on channel and spatial attention mechanism [4-8], the details can be seen in Fig 1 and Table 1. Our goal is to aggregate sufficient multi-scale contextual information and learn more effective feature representations.

2 Methodology

The proposed solution mainly includes two stages, each stage was formulated as a binary segmentation task by training a corresponding neural network. Specifically, we treat kidney and lesion as same category in the first stage, and the prediction output of the well-trained network are multiplied with input image to locate region of interest. In the second stage, these masked images are used to training another network with only lesion as foreground output. It is worth noting that we did not utilize any post processing. The entire pipeline is shown in Fig 2.

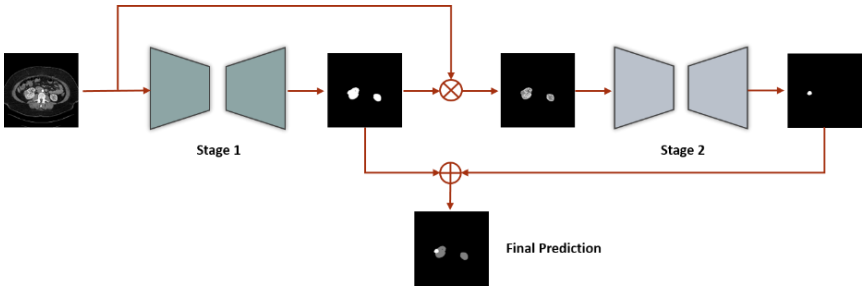

Fig. 2. The pipeline of the proposed solution for the kidney lesion segmentation. The final prediction is obtained by adding output masks from stage 1 and stage 2 together.

2.1 Data Preprocessing and Augmentation

We adopt 2D axial slices image extracted from 3D kidney CT scans as evaluation data, the preprocessing and data augmentation procedure consists of slices resampling, image scaling, cropping, and elastic transformation. All CT scans' intensity values are truncated to [-150, 250] to exclude irrelevant organs and tissues. We resample slices from each CT volume and save the images as local file. During training, random data augmentation was performed on the fly.

***Resampling*:** Since only a small part of slices within a CT volume contains kidney and tumor, there will be significant data imbalance problem along with huge computational burden if extracting all the slices (~40754) for the model training. Alternatively, for each kidney CT scan data, we extract all positive slices (slices contain the object content) while select the remaining negative slices (background slices) under a sampling interval of 5 (~23887).

***Scaling and cropping*:** Before fed into network for training, each image was rescaled with a random factor between 0.75 and 1.2, cropping and padding operation was then accordingly used to adjusting the image to the required size of model input (i.e. 256 or 384).

Elastic transformation: For further improving the robustness of the training process, we adopt elastic transformation in our data augmentation. To create an image

deformation, displacement field was generated first and then convolved with a Gaussian of standard deviation σ. We set the scale factor of the displacement field as 3, and σ =1 in this study.

2.2 Network Architecture

To learn patterns for capturing large object, output CNN features should correspond to sufficiently large receptive fields. On the other hand, for capturing small sized objects, output features should correspond to sufficiently small receptive fields to localize small regions of interest precisely. Follow this spirit, we re-design the components of U-net to aggregate multi-scale contextual information and improve the features representational ability.

Encoder path: In our model, the encoder path was organized in three consecutively stacked convolutional layers followed by alternatively layered basic encoder blocks and transition down blocks (down-sample layers). The first three convolutional layers are general convolution operation, while all other blocks using separable convolution. Each encoder block includes several densely connected parallel dilated convolutional (PDC) modules and a recalibration module. We set different dilation rate for different branch of PDC module, thus fruitful contextual information can be obtained by each encoder block. For further improvements, the recalibration module are used to re-weighting the feature maps with the guidance of descriptors aggregated from spatial dimension and channel dimension of the feature maps. Since pooling operation will cause information loss, hence the down-sample layer was constructed with a batch normalization layer and a separable convolutional layer.

In the task of kidney segmentation, all the strides of the down-sample layers were fixed as 2, the spatial resolution of the encoded feature maps is thus 16 times smaller than input image size. For the purpose of capturing small lesion structures as possible, we set stride as 1 for the last down-sample layer of the model in the lesion segmentation so as to maintain a relative large feature map size.

Decoder path: The decoder path mainly used to recover spatial resolution of feature maps by gradually incorporating fine features from encode path. Same to encoder path, each decoder block also equipped with a recalibration model to enhance the feature ability. The transition up blocks (up-sample layer) utilize transposed convolutional layer to enlarge the feature maps size.

Table 1. Details of proposed network architecture.

Modules	Layers
Input conv block	3×3 conv, s $= 1$, p $= 1, 18$ 3×3 conv, s $= 1$, p $= 1, 18$ 3×3 conv, s $= 1$, p $= 1, 36$
Encoder block 1	$\begin{bmatrix} 3 \times 3 \text{ sep conv, d} = (1,3,5), \text{p} = (1,3,5) \\ 3 \times 3 \text{ sep conv, s} = 1, \text{p} = 1 \end{bmatrix} \times 2, 84$
Transition down 1	3x3 sep conv, s=2, p=1, 84
Encoder block 2	$\begin{bmatrix} 3 \times 3 \text{ sep conv, d} = (1,3,5), \text{p} = (1,3,5) \\ 3 \times 3 \text{ sep conv, s} = 1, \text{p} = 1 \end{bmatrix} \times 3, 156$
Transition down 2	3x3 sep conv, s=2, p=1, 156
Encoder block 3	$\begin{bmatrix} 3 \times 3 \text{ sep conv, d} = (1,3,5), \text{p} = (1,3,5) \\ 3 \times 3 \text{ sep conv, s} = 1, \text{p} = 1 \end{bmatrix} \times 4, 252$
Transition down 3	3x3 sep conv, s=2, p=1, 252
Encoder block 4	$\begin{bmatrix} 3 \times 3 \text{ sep conv, d} = (1,3,5), \text{p} = (1,3,5) \\ 3 \times 3 \text{ sep conv, s} = 1, \text{p} = 1 \end{bmatrix} \times 4, 348$
Transition down 4	3x3 sep conv, s=2, p=1, 384
Encoder block 4	$\begin{bmatrix} 3 \times 3 \text{ sep conv, d} = (1,3,5), \text{p} = (1,3,5) \\ 3 \times 3 \text{ sep conv, s} = 1, \text{p} = 1 \end{bmatrix} \times 3, 420$
Transition up 1	3×3 transpose conv, s $= 2$, p $= 1, 348$
Decoder block 1	$[3 \times 3$ sep conv, s $= 1$, p $= 1] \times 2, 348$
Transition up 2	3×3 transpose conv, s $= 2$, p $= 1, 252$
Decoder block 2	$[3 \times 3$ sep conv, s $= 1$, p $= 1] \times 2, 252$
Transition up 3	3×3 transpose conv, s $= 2$, p $= 1, 156$
Decoder block 3	$[3 \times 3$ sep conv, s $= 1$, p $= 1] \times 2, 156$
Transition up 4	3×3 transpose conv, s $= 2$, p $= 1, 84$
Decoder block 4	$[3 \times 3$ sep conv, s $= 1$, p $= 1] \times 2, 84$
Output conv block	3×3 conv, s $= 1$, p $= 1, 84$ 3×3 conv, s $= 1$, p $= 1, 18$ 3×3 conv, s $= 1$, p $= 1, 1$
Sigmoid	

2.3 Training Procedure

We train our networks with a combination of dice and binary cross-entropy loss:

$$\mathcal{L}_{total} = \alpha\mathcal{L}_{Dice} + \beta\mathcal{L}_{Bce} \quad (1)$$

α fixed as 0.8 for both two training stage, and β set as 0.4 and 0.04 for kidney and lesion segmentation respectively.

Kidney segmentation*:* The aim of this step is to locate kidney region, and as aforementioned, kidney and tumor labels are merged as same category. We train the model from scratch using Adam optimizer with learning rate of 0.0003 and iterative epoch of 100.

Lesion segmentation: Once the model was well trained in the first stage, we obtain the prediction masks of kidney from the output score maps with threshold of 0.4. Subsequently, each image is multiplied with the corresponding prediction kidney mask, as shown in Fig 2, irrelevant objects are thus excluded which reduce the difficulty of the following lesion segmentation. In the second stage, we fine-tune the model from kidney segmentation using those masked images. The learning rate set as 0.0001 and training epoch fixed as 80.

After both models converged, final prediction masks of kidney and lesion was generated by summing score maps from the two models. The threshold fixed also as 0.4. In the submission phrase of the challenge, we take ensemble strategy since we have 10 models totally for the five folds training.

3 Experiment Results

3.1 Dataset and Implementation setting

The proposed model was trained and evaluated on the kidney CT data from 2019 Kidney Tumor Segmentation Challenge[1]. There are total 210 labeled CT scans released as training and validation data. Without using any external data, we perform five-fold cross validation using dice similar coefficient as measurement metric.

All the experiments are implemented based on python environment and Pytorch platform with a workstation of two Titan X GPUs. The source code can be found at http:// github.com.

3.2 Results

The evaluation results of kidney and lesion segmentation was shown in Table 2 and Table 3 respectively. Due to some prediction masks or ground truth masks contain no kidney or lesion objects, in this case, the way of calculating the Dice can be different. In our study, when both prediction and ground truth are all zeroes

[1] https://kits19.grand-challenge.org/home/

values, we compute the Dice as 1. However, when ground truth contain no positive categories, and prediction is not empty, the Dice is regarded as 0.

Table 1. Results on kidney segmentation.

Fold numbers	Dice coefficients	Mean IoU
Fold-1	96.18	94.22
Fold-2	93.96	91.92
Fold-3	93.86	91.92
Fold-4	94.55	92.58
Fold-5	94.47	92.44

Table 2. Results on lesion segmentation.

Fold numbers	Dice coefficients	Mean IoU
Fold-1	85.50	83.60
Fold-2	82.31	80.27
Fold-3	84.64	82.46
Fold-4	81.42	80.17
Fold-5	80.85	78.63

References

1. Heller, N., et al.: The KiTS19 Challenge Data: 300 Kidney Tumor Cases with Clinical Context, CT Semantic Segmentations, and Surgical Outcomes. arXiv preprint: 1904.00445 (2019).
2. Fisher, Y., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv: 1511.07122. (2015).
3. Ronneberger, O., Philipp, F., Thomas B.: U-net: Convolutional networks for biomedical image segmentation." International Conference on Medical image computing and computer-assisted intervention, pp. 234-241. Springer, Cham, (2015).
4. Huang, G., et al.: Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700-4708. (2017).
5. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1251-1258. (2017).
6. Liang-Chieh, C., et al.: Encoder-decoder with atrous separable convolution for semantic image segmentation. Proceedings of the European conference on computer vision, pp. 801-818. (2018).
7. Jie, H., Shen, L., Sun, G.: Squeeze-and-excitation networks. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7132-7141. (2018).
8. Roy, A. G., Navab, N., Wachinger, C.: Recalibrating Fully Convolutional Networks With Spatial and Channel “Squeeze and Excitation” Blocks. In IEEE Transactions on Medical Imaging, vol. 38, no. 2, pp. 540-549. (2019).