Automatic Kidney and Tumor Segmentation with Attention-based V-Net

Yucheng Hu, Han Deng, Yang Zhou, Yimin Chen, Hao Zhou, and Wanqi Yang
yc6820@outlook.com
yangwq@njnu.edu.cn

School of Computer Science and Technology, Nanjing Normal University

Abstract. Deep learning, especially Convolutional Neural Networks (CNNs) have been implemented to resolve a variety of both computer vision and medical image analysis problems recently. Among a rather wide range of Segmentation CNNs, V-Net is a relatively popular one, which is also an extended version of U-Net which processes 2D images. In this work, we propose an innovative V-Net with a embeded attention module. Inspired by spatial neural attention for generating pseudo-annotations, we modify the Decoupled attention into 3D version and insert it into the V-Net. This CNN network is trained end-to-end on CT volumes, and able to learn to predict segmentation blocks for a certain case. The definition of "block" will be elaborated in Section 3. Finally, the blocks will be concatenated to create a complete segmentation for a single case.

Keywords: Kidney and Tumor Segmentation · V-Net · Spatial Attention.

1 Introduction and Related work

There were more than 400,000 kidney tumor diagnoses all over the world in 2018 resulting in more than 175,000 deaths , up from 208,000 diagnoses and 102,000 deaths in 2002 [1]. With kidney tumor diagnoses increasing and the desire to do less impairment to patients, finding a novel and effective method to localize kidney tumor is rather crucial. When tumor details are precisely quantified, kidney cancer treatment decision (notably, the decision of Radical Nephrectomy(RN) and Partial Nephrectomy(PN)) is able to be made with more ease.

Although deep convolutional neural networks (CNNs) based methods have been widely used for segmenting human organs, most CNNs doesn't have universalities to be implemented on tumors segmentation. Due to the unique properties of extremely small sizes and low contrast, a majority of semantic segmentation networks are not capable to detect tiny structure of tumors. Therefore, revise conventional CNNs to make it adaptive to handle segmenting sophisticated objects is prevailing in medical imaging analysis.

Recently, attention models are introduced into CNNs in order to emulate humans' visual attention mechanism of focusing on essential parts for recognizing

objects in visual scenes. Mnih *et al*. [5] proposed a recurrent attention with hard alignment. Training hard attention, however, is rather difficult. As a result, some soft alignment attention-based models were also developed subsequently. For instance, Li *et al*. [3] utilize attention module to attend global contextual information to guide object detection. Having evaluated pragmatic features of attention models, we decided to implement attention mechanism to our semantic segmentation network.

2 Data

2.1 Data description

The Kits2019 challenge data are collected from 300 patients who underwent nephrectomy for kidney tumors. 210 cases of these patients are selected as training set, while 90 cases are selected as test set [1]. Those CT images usually have a height of 512 and a width with 512. The numbers of slices, however, varies from 32 to 1059, which would be tricky to handle preprocessing. Each case has 3 labels, which are background, kidney and kidney tumor respectively.

2.2 Data preprocessing

Firstly, we resized every slice into 256x256, which is one fourth of the original slice. Having considered the efficacy of our CNN model, which is related to the depth of the V-Net, we decided not to resize on the slice dimension. This is important for several downsampling layers in our model. Secondly, we cut a volume into multi blocks with regard to its slices. Given that the smallest slices are 32 in the training set, which also happpen to be a exponential number of 2, the block's size is determined to be 32x256x256. Finally, normalization function is utilized across individuals. That's to say, we subtract the minimum value and dividing the difference of maximum and minimum on each case independently.

3 Methods

The section introduces our proposed method. Our proposed network consists of 2 components: a modified V-Net derived from the original one, and a attention module which is embedded at the lowest part of the V-Net in order to learn some certain useful regional information about kidneys and Tumors.

3.1 Overall Architecture

The modified V-Net just has a little difference with its primitive version [4]. We cut the number of common convolutional layers into half or two thirds in a bid to save memories when training with a large dataset, which has a big impact on computing efficiency. In addition, in V-Net Fig.1, left part of downsamplings allows the network to reduce the size of input and to increase the receptive

field of the features being computed in subsequent layers. The right part of the network extracts features and expands spatial support for low resolution features in a bid to gather and integrate important information. Considering the significant attributes of the 2 parts mentioned above, we decide to insert the attention module between the left and right part to highlight some regions which are influential to segment final labels but can easily be omitted in low resolution.

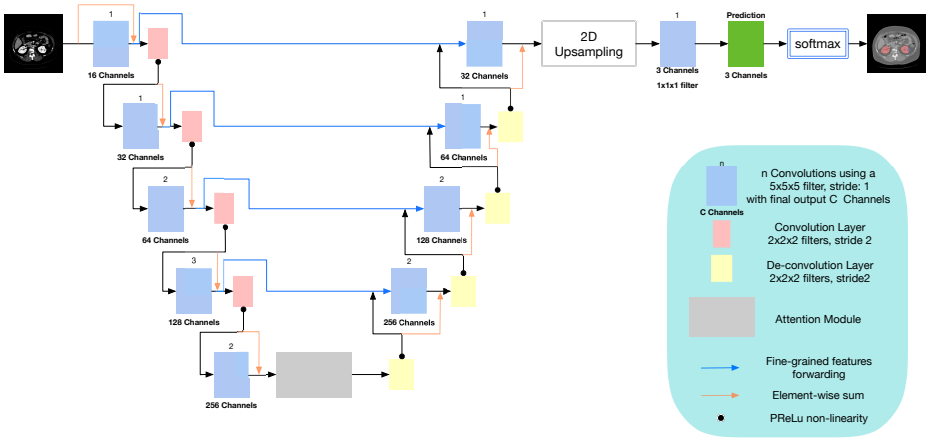

Fig. 1. Schematic representation of architecture of Attention-based V-Net. The left part downsamples features and expands receptive field, while the right part save high resoltion features and assemble them with low resolution ones.

3.2 Attention Module

The attention module is modified into 3D version from the Decoupled Spatial Neural Attention proposed by Zhang *et al*. [6] . Considering the conventional attentions' single interested region mining ability, the decoupled is split into 2 parts to alleviate this problem. In Fig.2, the upper branch in the figure is called Expansive attention detector, which aims at identifying object regions. The lower branch represent the Discriminative attention detector, which is intended to mine the discriminative parts. Such attention modules have some different attributes which are complementary to each other to combine features of global and local object regions for semantic segmentation.

3.3 Implementation Details

We use Adam solver [2] for training process with an initial learning rate of $1e-4$ and decays according to a specific function: $lr = lr_{init} \cdot 0.985^{epoch}$, as well as an l2 weight decay of $1e-5$. We refer to an epoch as an iteration over the whole training set (210 cases) with a batch size of 2 blocks.

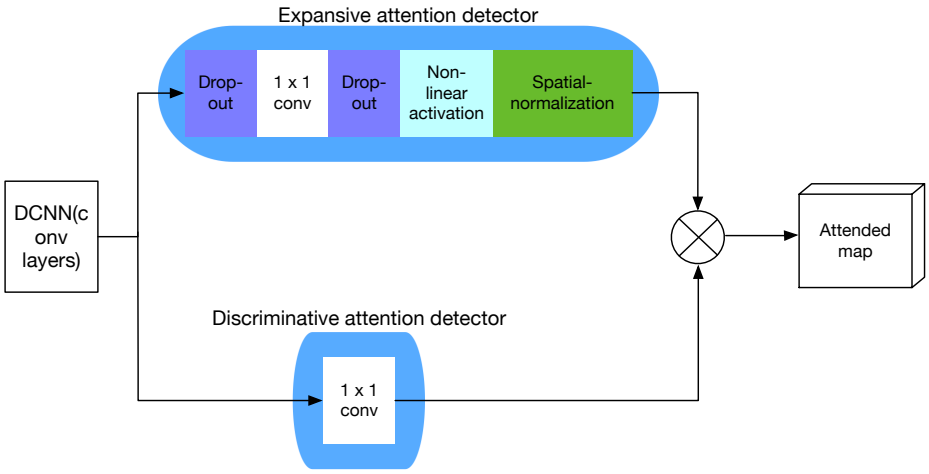

Fig. 2. Illustration of Decoupled Attention. The feature maps generated by it successfully integrate distinctive and contextual information.

References

1. Nicholas Heller, Niranjan Sathianathen, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, et al. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445*, 2019.
2. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
3. Jianan Li, Yunchao Wei, Xiaodan Liang, Jian Dong, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. Attentive contexts for object detection. *IEEE Transactions on Multimedia*, 19(5):944–954, 2016.
4. Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571. IEEE, 2016.
5. Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212, 2014.
6. Tianyi Zhang, Guosheng Lin, Jianfei Cai, Tong Shen, Chunhua Shen, and Alex C Kot. Decoupled spatial neural attention for weakly supervised semantic segmentation. *IEEE Transactions on Multimedia*, 2019.