Minimal Information Loss Attention U-Net for abdominal CT of kidney cancers segmentation

Yubai Liu Fucang Jia Shouliang Qi

Shenzhen Institutes of Advanced Technology Chinese Academy of Science

fc.jia@siat.ac.cn

Abstract. Recent work has shown that U-net is a straight-forward and successful architecture, it quickly evolved to a commonly used benchmark in medical image segmentation, Which nnU-Net had better performance We improved the nnU-Net model by incorporating a new image pyramid to preserve contextual features and attention gate. In order to let different kinds of class details more easily accessible at different scales, we injected the encoder layers with an input image pyramid before each of the max-pooling layers. We proposed a new image pyramid mechanism with dilated convolution that counters the loss of information caused by max-pooling, re-introducing the original image at multiple points within the network. We evaluated this model in the 2019 Kidney Tumor Segmentation Challenge. and got the dice coefficient 0.958 of kidney and 0.847 of tumors.

Keywords: Semantic Segmentation, Multiple input, Medical Imaging, U-Net

1 Introduction

There are more than 400,000 new cases of kidney cancer each year [1], and surgery is its most common treatment [2]. Due to the wide variety in kidney and kidney tumor morphology, people are currently concerned with how tumor morphology is related to surgical results, [3,4] and developing advanced surgical planning techniques [5]. Automatic semantic segmentation is a promising tool for these efforts, but morphological heterogeneity makes it a challenge.

Medical image analysis faces difficulty balancing precision and recall due to small regions-of-interest (ROI) found in medical images. Research efforts to address small ROI segmentation propose more discriminative models such as attention gated networks [6]. CNNs with attention gates (AGs) focus on the target region, with respect to the classification goal, and can be trained end-to-end. At test time, these gates generate soft region proposals to highlight salient ROI features and suppress feature activations by irrelevant regions. To address the issues of kidney cancer segmentation, we combine attention gated U-Net with a new image pyramid mechanism.

Our major contributions include: (1) a deeply supervised attention U-Net [5], improved with a multi-scaled input image pyramid for better intermediate feature representations. (2) a new image pyramid mechanism with dilated convolution.

2 METHODOLOGY

2.1 Network architectures

In this paper, we present the framework bases on the original U-Net [6] and the nnU-Net [8]. At the deepest stage of encoding, the network has the richest possible feature representation. However, with cascaded convolutions and non-linearities, spatial details tend to get lost in the high-level output maps. This makes it difficult to reduce false detections for small objects that show large shape variability [5]. To address this issue, we use soft attention gates (AGs)(Fig.3) to identify relevant spatial information from low-level feature maps and propagate it to the decoding stage.

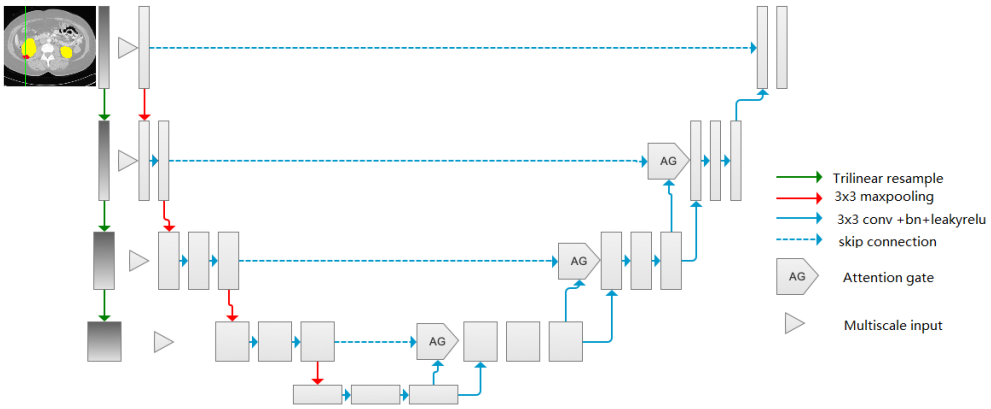

Fig. 1. Proposed Attention U-Net architecture with input image pyramid and deep supervised output layers.

Moreover, since different kinds of class details are more easily accessible at different scales, we inject the encoder layers with an input image pyramid (Fig.2) before each of the max-pooling layers. Combined with deep supervision, this method improves segmentation accuracy for datasets where small ROI features can get lost in cascading convolutions and facilitates the network learning more locality aware features with respect to the classification goal

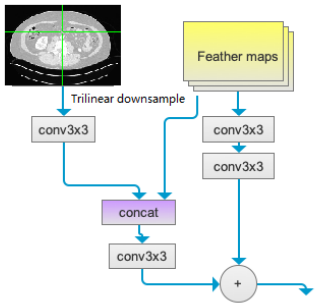

Fig. 2. A new image pyramid mechanism with dilated convolution. First, the image is sampled to the same resolution as the current layer with trilinear interpolation. Then after a 3x3 dilated conv. After concat with feather maps, add to feather maps.

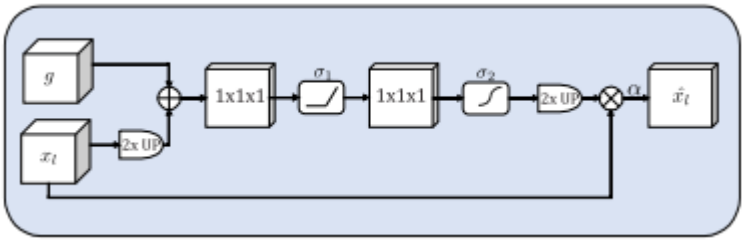

Fig.3. Schematic of additive attention gate (AG) adapted from [7]. Input features x_l are scaled with attention coefficients α_i to propagate relevant features to the decoding layer output $\hat{x}_l$. The coarser gating signal g provides contextual information while spatial regions from the input x_l provide locality information.

2.2 Experiment

Preprocessing. All data is cropped to the region of nonzero values. CNNs do not natively understand voxel spacings. In medical images, it is common for different scanners or different acquisition protocols to result in datasets with heterogeneous voxel spacings. To enable our networks to properly learn spatial semantics, all patients are resampled to the median voxel spacing of their respective dataset, where third order spline interpolation is used for image data and nearest neighbor interpolation for the corresponding segmentation mask. The described scheme is independently applied to each case and each modality. Only two levels of headings should be numbered. Lower level headings remain unnumbered; they are formatted as run-in headings. [8]

Training Procedure. Our U-Net uses 30 feature maps at the highest resolution layers. Here we start with a batch size of 2.

All models are trained from scratch and evaluated using five-fold cross-validation on the training set. We train our networks with a combination of dice and cross-entropy loss:

$$L_{total}=L_{dice}+L_{C\,E} \quad (1)$$

we compute the dice loss foreach sample in the batch and average over the batch. The dice loss formulation used here is a multi-class adaptation of the variant proposed in [9]. Based on past experience [10] we favor this formulation over other variants [11]. The dice loss is implemented as follows:

$$L_{dice}=-\frac{2}{|K|}\sum_{k\in K}\frac{\sum_{i\in I} u_i^k v_i^k}{\sum_{i\in I} u_i^k+\sum_{i\in I} v_i^k} \quad (2)$$

where u is the softmax output of the network and v is a one hot encoding of the ground truth segmentation map. Both u and v have shape I × K with i ∈ I being the number of pixels in the training patch/batch and k ∈ K being the classes. We use the Adam optimizer with an initial learning rate of 3×10^{-4} . We define an epoch as the iteration over 250 training batches. During training, we keep an exponential moving average of the validation(l_{MA}^{v}) and training (l_{MA}^{t}) losses. Whenever l_{MA}^{t} did not improve by at least 5×10^{-3} within the last 30 epochs, the learning rate was reduced by factor 5. The training was terminated automatically if l v MA did not improve by more than 5 $\times 10^{-3}$ within the last 60 epochs, but not before the learning rate was smaller than 10^{-6}.(8)

3 Results

Table 1. Results

	Dice
Kidney	0.958
Tumor	0.847

For the test cases we use the five networks obtained from our training set cross-validation as an ensemble to further increase the robustness of our models.

References

1. "Kidney Cancer Statistics." World Cancer Research Fund, 12 Sept. 2018, www.wcrf.org/dietandcancer/cancer-trends/kidney-cancer-statistics.
2. "Cancer Diagnosis and Treatment Statistics." Stages | Mesothelioma | Cancer Research UK, 26 Oct. 2017, www.cancerresearchuk.org/health-professional/cancer-statistics/diagnosis-and-treatment.
3. Kutikov, Alexander, and Robert G. Uzzo. "The RENAL nephrometry score: a comprehensive standardized system for quantitating renal tumor size, location and depth." The Journal of urology 182.3 (2009): 844-853.
4. Ficarra, Vincenzo, et al. "Preoperative aspects and dimensions used for an anatomical (PADUA) classification of renal tumors in patients who are candidates for nephron-sparing surgery." European urology 56.5 (2009): 786-793.

5. Taha, Ahmed, et al. "Kid-Net: Convolution Networks for Kidney Vessels Segmentation from CT-Volumes." arXiv preprint arXiv:1806.06769 (2018).
6. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in MICCAI. Springer, 2015, pp. 234–241.
7. Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al., "Attention u-net: Learning where to look for the pancreas," arXiv preprint arXiv:1804.03999,(2018).
8. F Isensee,et al." nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation." arXiv preprint arXiv:1809.10486.
9. M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," in Deep Learning and Data Labeling for Medical Applications. Springer, 2016, pp. 179–187.
10. F. Isensee, P. F. Jaeger, P. M. Full, I. Wolf, S. Engelhardt, and K. H. Maier-Hein, "Automatic cardiac disease assessment on cine-mri via time-series segmentation and domain specific features," in International Workshop on Statistical Atlases and Computational Models of the Heart. Springer, 2017, pp. 120–129.
11. C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. J. Cardoso, "Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations," in Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, 2017, pp. 240–248.