Paper Reading No.17

Deep Q Learning Driven CT Pancreas Segmentation with Geometry-Aware U-Net

Sheng Lian

October 2019

1 Brief Paper Intro

- ***Paper ref:*** TMI 2019 ,
  - http://arxiv.org/abs/1904.09120
  - https://doi.org/10.1109/TMI.2019.2911588
- ***Authors:*** Yunze Man, Yangsibo Huang, Junyi Feng, Xi Li, Fei Wu from Zhejiang University
- ***Paper summary:*** This paper proposed a Deep Q Network(DQN) driven approach with deformable U-Net to accurately segment the pancreas. Here, DQN based model learns a context-adaptive localization policy to produce a bounding box for finer segmentation. Then, deformable U-Net is used for fine segmentation.

- ***Reading motivation:*** Nowadays it seems that in the task of medical image segmentation, a coarse-to-fine structure is necessary for better performance. After reading some RL-related papers, I think DQN is a suitable way for the first step. Unfortunately (and inevitably), I found this paper.

2 Methodology

Medical image segmentation tasks like pancreas segmentation encounters numerous difficulties with small-sample-sized training, severe class imbalance, and background clutter with confusing distractions. This paper propose an anisotropic geometry-aware two-stage deformable deep learning scheme with a contextual interaction mechanism governed by a deep reinforcement learning (DRL) strategy. The overall framework can be summarized as:

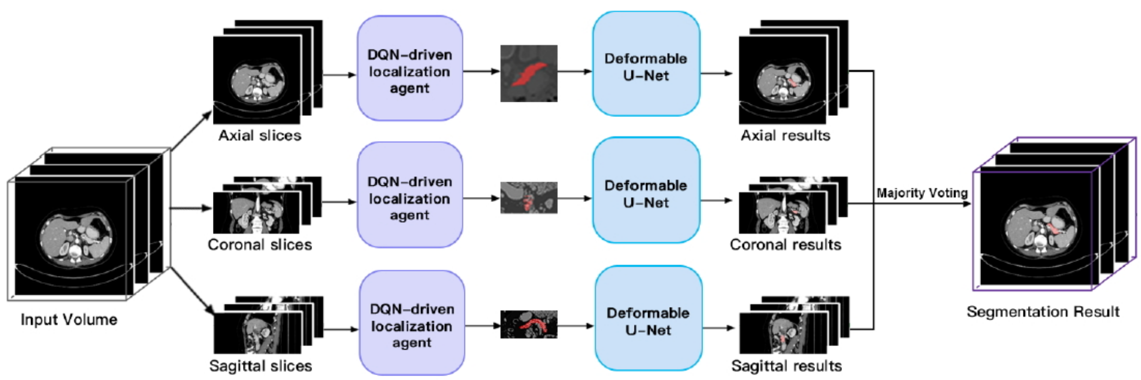

Figure 1: Overall architecture of the approach. A 3D volume is splitted into 2D slices of three axes.

$$F(X) = \text{Seg}(\text{Loc}(X)) \tag{1}$$

Here, the ***Loc()*** function is achieved by DQN, and the ***Seg()*** function is achieved by the proposed modified U-Net. The overall architecture of the approach is shown in Fig 1. The original 3D CT volume is sliced from three axis: sagittal, coronal and axial views. For each view, a view-specific network is trained with same structure. Each slice goes through Loc and Seg processes and obtain its segmentation result. Thepixel-wise results of three axes are merged by major voting to formulate the final voxel-wise mask of a 3D volume.

2.1 DRL-driven localization

This localization procedure is modelled as a Markov Decision Process(MDP). The actions are defined as Fig 2.

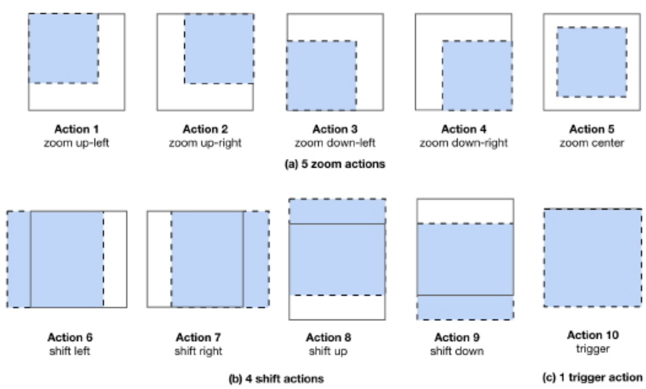

Figure 2: Illustration of (a)zoom, (b)shift and (c)trigger actions.

The agent will terminate in two conditions, either when the trigger action is selected, or when the agent reaches the maximum search step. The agent use modified IOU function as the quality measurement, which goes as

$$r_m(s, a) = \operatorname{sign}\left(\operatorname{IoU}\left(w', g\right) - \operatorname{IoU}(w, g)\right) \tag{2}$$

So, if the IOU is greater after taking an action, the reward is 1，and otherwise -1. Simple but effective.

For the trigger action, this approach pay extra attention to the recall of pancreas.

$$\mathrm{Recall}(w, g) = \mathrm{area}(w \cap g) / \mathrm{area}(g) \tag{3}$$

Here, g for the ground truth mask, and w for window. The trigger reward is positive only when:

$$r_t(s, a) = \begin{cases} +\sigma, & \mathrm{Recall}(w, g) > \tau_{\mathrm{Recall}}, \mathrm{IoU}(w, g) > \tau_{IoU} \\ -\sigma, & \text{Otherwise} \end{cases} \tag{4}$$

After these definition is the classical DQN related issues. Similar setting, including ϵ-greedy policy, experience replay, etc. are adopt here. Since these issues have been introduced in my former reading notes, I won't go into details here.

The DQN-driven localization is illustrated in Fig 3.

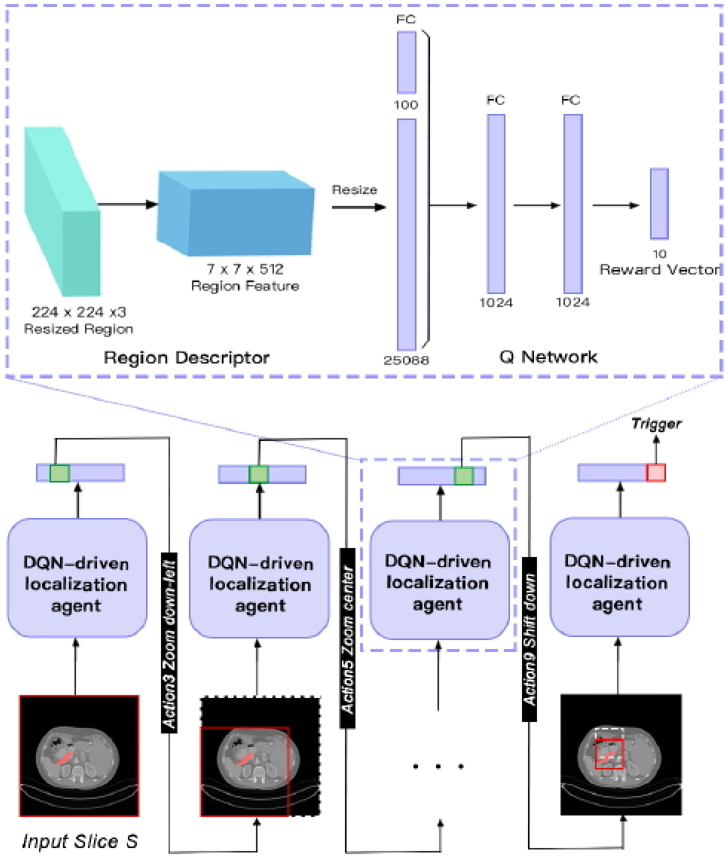

Figure 3: Illustration of DRL-driven localization.

2.2 Deformable Segmentation

The segmentation U-Net adopt Dice as loss function, which goes as

$$DL(F(X), Y) = -DSC(F(X), Y) = -\frac{2 \times |F(X) \cap Y|}{|F(X)| + |Y|} \qquad (5)$$

Specifically, in order to solve the problem of variable pancreas size, authors adopt deformable convolution. Deformable convolution [1] is a mechanism which allows receptive field to be learned rather than be fixed in standard convolution, which can be referred in the literature [1]. Generally speakinging, the offset map is obtained by adding a convolutional layer after the last feature map.

The majority voting indicated in Fig 1 is given by

$$Y = Majority\,(Y_s, Y_c, Y_a) = \left\lfloor \frac{1}{2} + \frac{Y_s + Y_c + Y_a}{3} \right\rfloor \tag{6}$$

In other words, a pixel is finally classified as pancreas only when at least two of the views take it as pancreas.

3 Conclusion

The proposed method achieved a significant improvement in performance, which can be referred in the paper, and I will not go into details here. Overall, this paper use DQN for localization and deformable U-Net for finer segmentation. There are not many innovations in the method, but the experiments are solid and this paper tells a very good and complete story.

References

[1] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.