DA-RefineNet: Dual-inputs Attention RefineNet for Whole Slide Image Segmentation
Ziqiang Li1, Rentuo Tao1, Qianrun Wu2, Bin Li1
1University of Science and Technology of China
2Hefei University of Technology
iceli@mail.ustc.edu.cn

Abstract
Automatic medical image segmentation has wide applications for disease diagnosis. However, it is much more challenging than natural optical image segmentation due to the high-resolution of medical images and the corresponding huge computation cost. The sliding window is a commonly used technique for whole slide image (WSI) segmentation; however, for these methods based on the sliding window, the main drawback is lacking global contextual information for supervision. In this paper, we propose a dual-inputs attention network (denoted as DA-RefineNet) for WSI segmentation, where both local fine-grained information and global coarse information can be efficiently utilized. Sufficient comparative experiments are conducted to evaluate the effectiveness of the proposed method; the results proved that the proposed method can achieve better performance on WSI segmentation compared to methods relying on single-input.

Motivation
Figure 1: Demonstration of training medical images and its corresponding segmentation results. (a) training samples in ICIAR2018 dataset, where green rectangles denote sliced images; (b) segmentation results of (a), color pixels represent different region types (normal, benign, in situ, invasive).

It’s easy to find that sliced patches with similar texture sometimes were labeled differently due to the context discrepancy, which can be seen in the left image of Fig.1. Intuitively, this may caused by the discrepancy between surrounding regions of different slice images. For a specific region, more informative context information can be derived by simply increasing the size of slice window; however, it will also bring more computation costs. In order to balance slice image size and global context information, we propose a dual-inputs attention network named DA-RefineNet to combine fine texture features and coarse spatial features together, which allows a larger receptive field with little computation cost increasing.

Method: Dual-input Attention RefineNet
Framework
The architecture of proposed DA-RefineNet was demonstrated in Fig. 2, where we can see the model mainly composed of three parts: two encoders (ENC Sabha ENC Saball) for slice image and full image processing, and a refine decoder (DEC Sabha DEC Saball) for producing segmentation results. Color arrows in green, blue, orange and gray represent down-sampling, skip-connection, up-sampling and feature processing, and a refine decoder (DEC Sabha DEC Saball) for producing segmentation results. Color arrows in green, blue, orange and gray represent down-sampling, skip-connection, up-sampling and feature fusion operations respectively.

Attn-Refine Block and Feature Fusion
Based on the intuition that global coarse images along with fine-grained local images features can help improve model performance, we adopt attention mechanism and proposed the attention refinement block(Attn-Refine). Detailed architecture of the proposed Attn-Refine block can be seen in the top part of Fig. 2, where color arrows represent different type of input features. Attn-Refine block contains three individual components: Attention block, RCU (residual convolution unit) and CRP (chain residual pooling). The proposed attention block is designed for feature fusion, which can be seen in the bottom part of Fig. 2. To make comparison, we keep RCU and CRP module the same as RefineNet [1].

Figure 2: Attn-Refine Block: the top part denote the architecture of a attention refine block while the bottom part denote the attention block structure.

Experiments

<table>
<thead>
<tr>
<th>Dual-inputs Encoder</th>
<th>Fusion strategy</th>
<th>[MIoU 1, IoU 1, MIoU 2, IoU 2, Accuracy Score, Params]</th>
<th>Dual-inputs vs Single-input</th>
<th>Accuracy Score</th>
<th>Params</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet50 + Add</td>
<td>Concat</td>
<td>48.9 61.1 38.9 39.2 58.3 75.7 71.5</td>
<td>48.9 61.1 38.9 39.2 58.3 75.7 71.5</td>
<td>580M</td>
<td>517M</td>
</tr>
<tr>
<td>ResNet101 + Add</td>
<td>Concat</td>
<td>47.5 58.3 36.6 42.0 60.0 74.4 71.4 580M</td>
<td>47.5 58.3 36.6 42.0 60.0 74.4 71.4</td>
<td>580M</td>
<td>517M</td>
</tr>
</tbody>
</table>

Table 1: Quantitative Comparison of Dual-inputs vs Single-input

<table>
<thead>
<tr>
<th>Feature Fusion Strategies</th>
<th>Dual-inputs Encoder</th>
<th>Fusion strategy</th>
<th>[MIoU 1, IoU 1, MIoU 2, IoU 2, Accuracy Score]</th>
<th>Single-input</th>
<th>Accuracy Score</th>
<th>Params</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet50 + Add</td>
<td>Concat</td>
<td>48.9 61.1 38.9 39.2 58.3 75.7 71.5</td>
<td>48.9 61.1 38.9 39.2 58.3 75.7 71.5</td>
<td>580M</td>
<td>517M</td>
<td></td>
</tr>
<tr>
<td>ResNet101 + Add</td>
<td>Concat</td>
<td>47.5 58.3 36.6 42.0 60.0 74.4 71.4 580M</td>
<td>47.5 58.3 36.6 42.0 60.0 74.4 71.4</td>
<td>580M</td>
<td>517M</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Quantitative Comparison of Feature Fusion Strategies

Conclusions
• We proposed a new attention based dual-inputs framework for whole slide image segmentation, which can incorporate global image information and obtain larger receptive fields. It proves that the proposed method can achieve better performance compared to methods that rely on single-input.
• We explore several feature fusion strategies and propose a simple but effective fusion strategy based on attention mechanism under the intuition that coarse global features can help reorganize fine-grained local features.

References

Acknowledgements
This work was supported by the National Natural Science Foundation of China under grant No.19128044 and No.61386011. We also want to thank the data provider organizer of the ICIAR2018 Grand Challenge.