U-shaped Feature Extractor Used on Mask R-CNN for Cell Nuclei Image Segmentation

Qing Zhu¹, Xiaoyang Huang¹, Liansheng Wang¹* and Jonathan Li¹²

¹Fujian Key Laboratory of Sensing and Computing for Smart Cities, School of Informatics, Xiamen University, Xiamen, FJ 361005, China
²Department of Geography and Environmental Management and Department of Systems Design Engineering, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada
Email: lswang@xmu.edu.cn

Abstract. The cell nuclei segmentation of is a challenging task in microscopy image analysis. The problems of noise, small cell nuclei, and few training data samples in the data set will all affect the effectiveness of the model to varying degrees. This paper presents a new approach to nuclear image segmentation based on convolutional neural networks. Our approach is based on Mask R-CNN with some modification, which combines low-level semantic features for model training. In order to make the model specific to each low-level feature map, the attention mechanism was used to assign weights to each low-level feature map, making the model learning more purposeful. Our method achieves an average precision value of 62.8%, which is 2.7% higher than that of the Mask R-CNN (ResNet50) basic model and 6.3% of the Mask R-CNN (ResNet101) basic model.

1. Introduction

With the development of deep learning, Convolutional Neural Network (CNN) [1] has made significant progress in image segmentation, classification, detection, and other fields, and CNN has been proven to be significantly better than traditional computer vision methods. Most of the cells in the human body contain a nucleus full of DNA and automate nucleus detection, and segmentation is an important task. It can not only help doctors make a correct analysis of lesions faster, but also shorten the duration of drug testing. Due to the complex structure of the nucleus, a picture often contains many nuclei. Doctors may only make misdiagnosis with the naked eye. Therefore, it is significant to research the detection and segmentation methods of the nucleus.

In recent years, people have conducted much research on image detection and segmentation, and have made significant progress. The calculation efficiency based on the traditional convolutional neural network (CNN) method is low, and the effect is not ideal. A fully convolutional network (FCN) was proposed in [2] to solve this problem. The input of the network can accept images of any size, but the connection between pixels is not fully considered, resulting in insufficient segmentation accuracy. An encode-decode structure, which uses a jump connection, was proposed in [3] to allow the network to propagate context information to a higher level. This model has been proven to have excellent performance in the segmentation of the boundary position. The network designed in [4] used a feature extractor to input the extracted features into the second CNN network to achieve a better segmentation effect, but this method is not an end-to-end network. The Atrous Spatial Pyramid Pooling was used in [5] to capture multi-scale information with parallel cavity convolutions of different sampling rates. An adaptive attention mechanism module was proposed in [6], which considered the importance of
different channels in the feature map and can adaptively assign corresponding weights to the feature map. A residual structure was proposed in [7] to possibly train a very deep network. A complex residual connection structure was proposed in [8] for each encoder and decoder. The deep learning method was applied in [9] to the target detection task. CNN is used to extract deep features in the generated candidate boxes, and then these features are sent to the SVM classification. The classifier judges the category, and finally uses the regressor to fine-tune the position of the candidate frame, but this is not an end-to-end network, and the speed is relatively slow. The structure proposed by [10] solves the problem of slow speed to a certain extent, but it still cannot meet real-time applications, and the accuracy is still low. [11] Proposed RoI Align structure makes RoI and proposal achieve position alignment, improve the accuracy of a classification and bounding box detection, but it is not very good at detecting small objects. [12] proposed a convolutional neural network combined with multiple views for segmentation, and achieved the best segmentation effect at that time. [13] The proposed model can combine historical information to learn historical sampling experience, so that the segmentation result achieves the desired effect. [14] Using the cascade architecture to improve the Mask information flow in Mask R-CNN makes the segmentation results more accurate.

Due to a large number of nuclei and the relatively small size, this paper applies a U-shaped structure to the feature extraction stage of Mask R-CNN, so that it can make full use of the low-level feature, thereby improving the detection accuracy of small objects. The rest of the paper is organized as follows: Section 2 introduces our method in detail, Section 3 introduces the experiment and the experimental results, and Section 4 summarizes the full text.

Figure 1. These two figures show the Mask R-CNN model and the feature extraction module we used

2. Method
Convolutional neural networks can have excellent feature extraction capabilities, and well-extracted features can often make the model have better performance. Mask R-CNN [11] uses ResNet50 or ResNet101 [7] as a feature extractor. Because it uses residual connection instead of directly learning the original features, the network layer can become deep without degradation, and thus has better performance. Because the low-level layer of the network contains the rich low-level feature, it is
beneficial to generate segmentation masks, and Mask R-CNN does not make full use of low-level features in the feature extraction stage, which leads to its inability to detect and segment small objects.

As shown in Figure 1(A), it is the framework of Mask R-CNN [11], and its feature extract module part adopts ResNet50 or ResNet101. Due to the use of the residual connection, this module can reach a very deep level, and achieve better feature extraction effect, but it does not make full use of low-level semantic information, which makes the segmentation and detection of small objects not very good.

Inspired by [3], we adopted a U-shaped network model for feature extraction, while still using the residual connection, which allows the model to fully utilize low-level features while still having a deeper level network structure, because each low-level feature contributes to the performance of the network is not the same, so we use SE Block[6] to give each low-level feature a corresponding weight. The network architecture is shown in Figure 1(B), which consists of a down-sampling path (left) and an up-sampling (right) path. In the down-sampling path, it is reused by MaxPooling and the convolution module, where Identity block uses a residual connection, and after each convolution, a Swish activation function [16] is adopted, as shown in Equation 1.

\[
f(x) = x \times \frac{x}{1 + e^{-x}}
\]  

where x is the input feature map.

In the up-sampling path, each step of up-sampling will copy the corresponding feature map in the down-sampling to the up-sampling by channel. Because the importance of the feature maps between different channels is different, we have adopted the attention mechanism SE block, and it can learn the feature weights through the loss of the network, making the effective feature maps have large weights, the invalid or small effect feature maps have small weights, so that the model can adaptively assign corresponding weights to each feature map. This is used to indicate the importance of different features.

3. Experiments and Results

We conduct the experiments on Nuclear datasets [15], the data set contains 670 training samples, the resolution size of each picture is 320*256, because the test set of this data set does not give a label, We divided the training set into 10 for evaluation. In addition, we have made three types of data augmentation:

- Flip each training sample randomly in both horizontal and vertical directions
- Gaussian blur is adopted for each training sample
- Color dither each training sample

In this section, the first part introduces the comparative experiment, and the second part introduces the ablation experiment.

3.1. Comparative Experiments

We conducted two sets of comparative tests. The results of the comparative tests are shown in Figure2 and Table1. In Figure2, we show the results of the two sets of comparative tests. From the Figure2, we can see that our method is better than the other two methods. More nuclei were detected, especially incomplete nuclei at the corners of the picture.

Our evaluation method uses the average precision of intersection over union (IoU) pixels under different thresholds. The IoU can be calculated by Equation 2.

\[
IoU(A,B) = \frac{A \cap B}{A \cup B}
\]  

where A is the box predicted by the model, and B is Groud Truth.

The average accuracy of pixels under each IoU threshold is calculated by Equation 3.

\[
Accuracy = \frac{TP(t)}{TP(t) + FP(t) + FN(t)}
\]
where \( t \) is the threshold, TP is the positive sample point of the correct classification, FP is the negative sample point of the incorrect classification, and FN is the positive sample point of the incorrect classification. The average accuracy under all thresholds as the average accuracy of the image is finally calculated.

As shown in Table 1, we calculate the average accuracy of each point of IoU under different thresholds. The thresholds are 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, and 0.95, respectively. Finally, we average the accuracy of these 10 different thresholds as the average accuracy of the image. As can be seen from the table, Mask R-CNN (ResNet101) performance is not as good as Mask R-CNN (ResNet50). This is because the size of the cell nucleus on the data set is small, and the deep semantic information is small helpful for detecting small objects. Our method uses the low-level feature, so it is better than the above two models in detecting small objects.

As shown in Figure 2, in this case, shows the comparison of the results of the basic model and our method.

### 3.2. Ablation Experiments

In the experiment, we did three sets of ablation experiments, as shown in Figure 3 and Table 2. In this summary, we call the model without SE Block and Swish activation function as the base model, and the model with SE Block as base+SE, and the model with SE Block and Swish activation function as base+SE+Swish.
Table 1. Accuracy of the experimental results of our method and other methods under different thresholds

| Threshold | 0.50 | 0.55 | 0.60 | 0.65 | 0.70 | 0.75 | 0.80 | 0.85 | 0.90 | 0.95 | Ave |
|-----------|------|------|------|------|------|------|------|------|------|------|-----|
| Mask R-CNN (ResNet50) | 0.832 | 0.829 | 0.826 | 0.813 | 0.801 | 0.781 | 0.767 | 0.739 | 0.059 | 0.000 | 0.601 |
| Mask R-CNN (ResNet101) | 0.838 | 0.830 | 0.816 | 0.800 | 0.766 | 0.699 | 0.574 | 0.284 | 0.045 | 0.000 | 0.565 |
| Ours | 0.898 | 0.880 | 0.845 | 0.825 | 0.797 | 0.777 | 0.704 | 0.459 | 0.099 | 0.000 | 0.628 |

Figure 3. This figure shows the accuracy of the experimental results of our method and other methods under different thresholds

Table 2. Accuracy of the experimental results of our method and other methods under different thresholds

| Threshold | 0.50 | 0.55 | 0.60 | 0.65 | 0.70 | 0.75 | 0.80 | 0.85 | 0.90 | 0.95 | Ave |
|-----------|------|------|------|------|------|------|------|------|------|------|-----|
| Base | 0.882 | 0.863 | 0.842 | 0.823 | 0.771 | 0.693 | 0.519 | 0.168 | 0.006 | 0.000 | 0.556 |
| Base+SE | 0.862 | 0.850 | 0.838 | 0.827 | 0.808 | 0.762 | 0.628 | 0.371 | 0.058 | 0.000 | 0.600 |
| Base+SE+Swish | 0.898 | 0.880 | 0.845 | 0.825 | 0.797 | 0.777 | 0.704 | 0.459 | 0.099 | 0.000 | 0.628 |

As shown in Figure 3 and Table 2, the accuracy of Base+SE model is higher than the Base model when the threshold is greater than 0.65, and the Base+SE+Swish model is better than the previous two
models, which shows that we use the SE Block And Swish activation function is effective for the detection task of the cell nucleus.

4. Conclusion
This paper proposes an improved model based on Mask R-CNN for the detection and segmentation of nuclei. Because Mask R-CNN does not make full use of low-level features, it is not good at detecting and segmenting nuclei. On this basis, this paper makes full use of low-level features, combined with the attention mechanism and Swish activation function, which makes the model have a certain improvement in the detection and segmentation of cell nuclei.

5. References
[1] A. Krizhevsky, I. Sutskever, G. E. Hinton, 2012. ImageNet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems, pp. 1097–1105
[2] J. Long, E. Shelhamer, and T. Darrell, 2015. Fully convolutional networks for semantic segmentation." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, DOI: 10.1109/CVPR.2015.7298965.
[3] O. Ronneberger, P. Fischer, and T. Brox 2017 "U-Net: Convolutional networks for biomedical image segmentation," in MICCAI, Munich, Germany
[4] A. Zhang et al. 2017 "Automated pixel-level pavement crack detection on 3D asphalt surfaces using a deep-learning network," Comput-Aided Civ. Inf., vol. 32, no. 10, pp. 805-819
[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille 2016. Deeplab: Semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFs. arXiv preprint arXiv:1606.00915
[6] J. Hu, L. Shen, and G. Sun, 2018. "Squeeze-and-excitation networks." Proceedings of the IEEE conference on computer vision and pattern recognition
[7] He, Kaiming, et al 2016 "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition
[8] Ibtehaz, Nabil, and M. Sohel Rahman 2020 "MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation." Neural Networks 121 p 74-87.
[9] Girshick, Ross, et al 2014 "Rich feature hierarchies for accurate object detection and semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition
[10] Ren, Shaoqing, et al. 2017 "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems. 2015.
[11] He, Kaiming, et al. 2017 "Mask r-cnn." Proceedings of the IEEE international conference on computer vision.
[12] Tokunaga, Hiroki, et al. 2019 "Adaptive weighting multi-field-of-view CNN for semantic segmentation in pathology." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
[13] Wang, Feigege, et al. 2019 "Context-aware Spatio-recurrent Curvilinear Structure Segmentation." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
[14] Chen, Kai, et al. 2019 "Hybrid task cascade for instance segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition.
[15] Kumar, Neeraj, et al. 2017 "A dataset and a technique for generalized nuclear segmentation for computational pathology." IEEE transactions on medical imaging 36.7 p 1550-1560.
[16] Ramachandran, Prajit, Barret Zoph, and Quoc V. Le 2017 "Swish: a self-gated activation function." arXiv preprint arXiv:1710.05941 7.