DLCV HW3 Report

111550098 楊宗儒

Introduction

1. Task

In this homework, we need to train a Mask-RCNN[1] based model to segment the instance in images of cells from multiple organs(See Figure 1). The dataset consists of about 300 images, and there are 4 classes of instances.

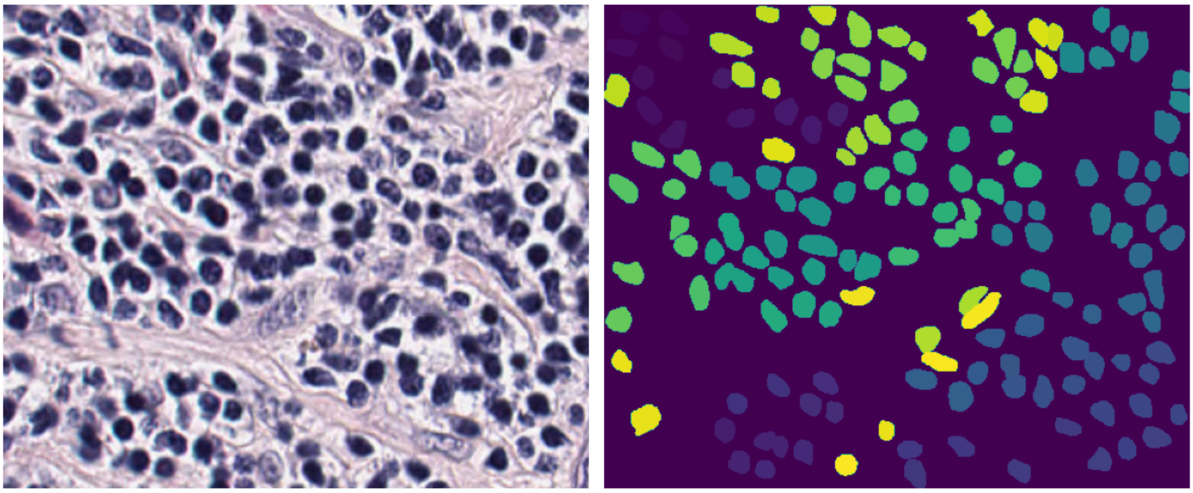

Figure 1. The sample of images in the dataset. The left is the input image, and the right is an instance map of one class.

2. Core Idea

The core idea of our approach is to utilize the pretrained weights of the feature pyramid networks [2] backbone, and adjust the anchor size for the Mask-RCNN[1] model to fit the size of the cell.

Method

1. Backbone

In our backbone, we use the feature pyramid network based on ResNet101 [3] to construct the pyramid.

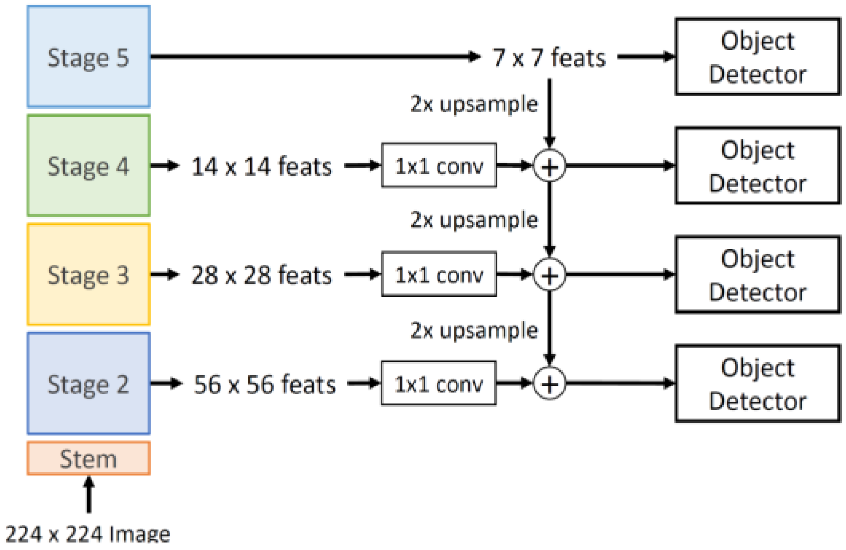

Figure 2. An Illustration of Feature Pyramid Network (from [4])

During training, the ResNet backbone is initialized to the pretrained weight that is trained on the ImageNet dataset [5].

2. **Anchor Generator**

   The input size of the image is 1024x1024, and our anchor generator uses the size (16, 32, 48, 64, 96) (pixels) and aspect ratios of (0.25, 0.5, 0.75, 1.0, 1.5, 2.0, 3.0, 4.0) for anchor. The design philosophy is to align the proposed region with the size of the cell.

3. **Hyperparameter**

   In the experiment, we initialize the weight as mentioned in the previous section, and we train the model for 150 epochs. The learning rate is 5e-5 with Adam optimizer, and the batch size is 2.

Result

The training curve and the performance of the model are in Figure 3. and Table 1.

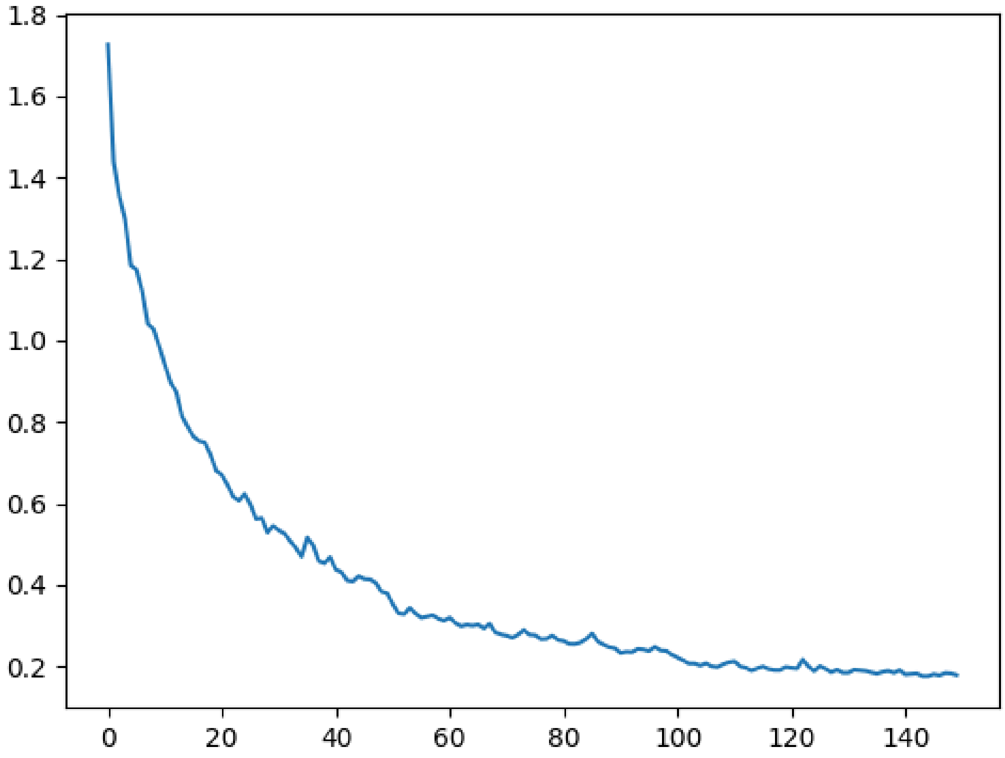

Figure 3. The Training Curve of the Model

Table 1. The Performance of Our model

Model	mAP50 Score
ResNet101_FPN(Ours)	0.357

Additional Experiment

1. Scaling Backbone

As a general rule in deep learning, scaling up a model often yields better performance. In this experiment we observe a similar result. However the difference is not as large, which might be due to the limited data amount providing less guidance on the backbone.

Table 2. The Performance of Different Backbones

Model	mAP50 Score
ResNet18_FPN	0.335
ResNet50_FPN	0.334
ResNet101_FPN(Ours)	**0.357**

2. Anchor Size

As a key component in Mask R-CNN[1] base model, the anchors serve as the candidates for possible instance regions. The size misalignment of anchors would result in poor instance localization and poor segmentation.

By default Pytorch uses the anchor sizes of (32, 64, 128, 256, 512) (pixels) and aspect ratios of (0.5, 1.0, 2). In our task it would be too large so we use (16, 32, 48, 64, 96) (pixels) and aspect ratios of (0.25, 0.5, 0.75, 1.0, 1.5, 2.0, 3.0, 4.0) for proximity to the size of our target instances(cells)

In this experiment we observed that such modifications give a better prediction.

Table 3. The Performance of Different Backbones with Similar Model Size

Model	mAP50 Score
Default Anchor Size	0.339
Ours	**0.357**

References

[1] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).

[2] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection.

In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2117-2125).
[3] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
[4] Slide of Chapter "Object Detection" in DLCV 2025
[5] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.

GitHub Link

<to be insert>