DLCV HW2 Report

111550098 楊宗儒

Introduction

1. Task

In this homework, the task is to detect the digits in some outdoor images of door plates and predict the entire number on them with Faster-RCNN [1] based models. The dataset consists of around 30,000 images of door plates for the training set and 13,000 for the testing set. (See Figure 1)

Figure 1. The sample of images in the dataset.

2. Core Idea

The core idea of our approach is to utilize the pretrained weights on faster-RCNN provided by pytorch for RPN and ROI-head and change the backbone to Feature Pyramid Network (FPN) [2] based on ResNet101 [3].

Method

1. Backbone

In our backbone, we use the feature pyramid network based on ResNet101 which takes the outputs of conv3_x, conv4_x and conv5_x layers (Referring to [resnet paper]) to construct the pyramid.

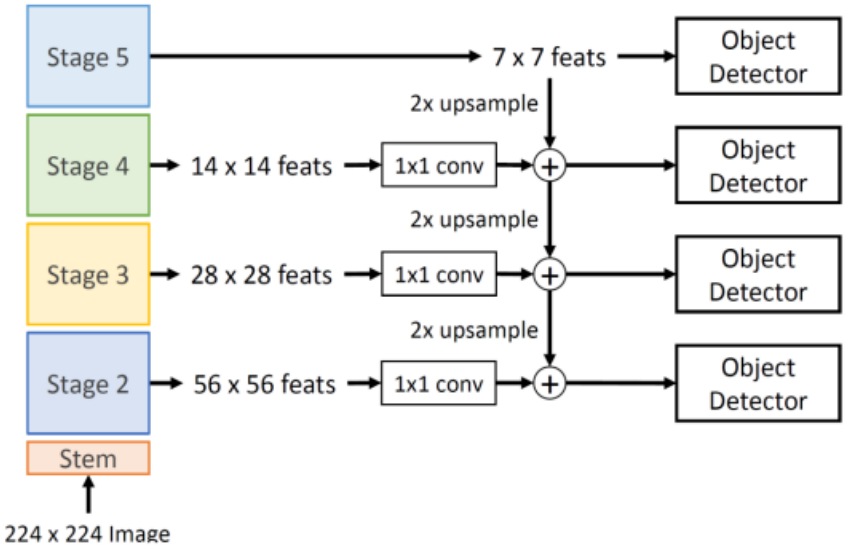

Figure 2. An Illustration of Feature Pyramid Network (from [6])

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

2. **RPN(neck)**

   The architecture of RPN is one layer of the Convolution-BatchNorm-ReLU layer.
   This design choice is to match the default setting of Faster-RCNN in Pytorch (fasterrcnn_resnet50_fpn), in which the pretrained weight is provided.

3. **Head**

   The classification head and the bounding box head are both single linear layers.
   Since the number of classes is different, this part of the network is trained from scratch.

4. **Hyperparameter**

   In the experiment, we initialize the weight as mentioned in the previous section, and we train the model for 6 epochs. The learning rate is 1e-4 with Adam optimizer, and the batch size is 4.

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	mAP Score	Accuracy
ResNet101_FPN(Ours)	0.36	0.75

Additional Experiment

1. Scaling Backbone

As the rule of thumb in deep learning, scaling up the model often gives a better performance.
In this task, we also see such relations in our experiments, especially when switching from ResNet50_FPN to ResNet101, the accuracy improves significantly.
The setting is the same as the previous section.

Table 2. The Performance of Different Backbones

Model	mAP Score	Accuracy
ResNet50_FPN	0.31	0.52
ResNet101_FPN(Ours)	**0.36**	**0.75**
ResNet152_FPN	**0.36**	0.73

2. ResNet vs ResNeXt

ResNeXt [4] claimed that with the same number of parameters, the performance of the model is better than ResNet.
However, in our experiment, switching the backbone to ResNeXt gives a similar performance with the ResNet.

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

Model	mAP Score	Accuracy
ResNeXt101_64x4d_FPN	**0.36**	0.74
ResNet101_FPN(Ours)	**0.36**	**0.75**

References

[1] Ren, S., He, K., Girshick, R., & Sun, J. (2016). Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence, 39(6), 1137-1149.
[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] Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1492-1500).
[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.
[6] Slide of Chapter “Object Detection” in DLCV 2025

GitHub Link

https://github.com/s0n9Yu/DLCV-hw2