Improvement Based on an Instance Segmentation Algorithm

Chen Mingyang

College of Computer Science and Technology, Qingdao University, Qingdao 266071, China.

cmylc123@163.com

Keywords: Segmentation Algorithm, Improvement

Abstract: In recent years, Mask-R-CNN has achieved great success in the field of image segmentation. This paper proposes an accurate image segmentation method based on Mask-R-CNN.We improve the boundary segmentation precision by modifying the mask branch structure. Used feature fusion on the mask branch to enhance the combination of semantic information and shallow information, and improve the segmentation effect. Expanded the size of the feature map of the RoIAlign layer, get more accurate location information. The experiment shows the accuracy of the algorithm.

1. Introduction

The instance segmentation is to mark different instances from the image using the object detection method, and then perform pixel-by-pixel tagging in different instance regions. Image segmentation is the basis for image understanding of computer vision, it is one of the most important steps in the image analysis process, and the segmented region can be used as the object for subsequent feature extraction. Currently, based on the deep convolutional neural network method, there are multiple instances of segmentation. Bharath et al. [1] proposed a super-column method, the activation tandem of all nodes of corresponding to the pixel network is used as a feature, perform synchronous detection and segmentation of object. He et al. Dai et al. [2] first used the regional proposal network to predict the uncategorized bounding box position and score. Using the first stage of the convolution feature and the bounding box as input, extracting features by ROIpooling, connect the two fully connected layers, the first fully connected layer is for dimensionality reduction and the second is for pixel level masking. Used bounding box and mask as input and finally output the classification score of each instance. Kaiming et al. [3] proposed a simple and effective example segmentation system. It is based on Faster R-CNN [4], and the base layer network selects the ResNet-101 [5] network. By using a multi-scale, multi-level pyramid structure to construct FPN (Feature Pyramid Network)[6], and adding a mask branch of a full convolution network based on FPN, added a small overhead, which has achieved good results in terms of segmentation accuracy and algorithm efficiency. The feature map obtained in the RoIAlign layer has a small resolution, which causes the loss of boundary pixel information, it has a great impact on applications that require high boundary segmentation accuracy. For this problem, this paper is based on the basic architecture of Mask-R-CNN, proposed a method to improve the prediction mask branch network structure and boundary refinement, and obtain better experimental results. We use the fusion of the front and back layer feature maps, the front layer can retain the shallow information, the latter layer can retain the deeper information, and the combination of the two further improves the boundary segmentation. Because the instance segmentation is for the pixel point operation, when performing the RoIAlign operation on the mask branch, the pixel points on the feature map should be retained as much as possible, the feature resolution is increased to 28×28 by bilinear interpolation, and finally obtained accurate boundary information.

  DOI: 10.25236/icscbd.2018.005

2. Approach

2.1 Introduction to Convolutional Neural Networks

As the convolutional network has achieved great success in instance segmentation, our framework is also based on CNN. The concept of convolutional neural networks originated from scientists' Receptive Field, each animal's neurons will only process a visual image of a small area, which is equivalent to the processing of the convolution kernel in CNN. Later, the concept of neurocognitive machine was proposed. The neurocognitive machine contains two types of neurons, one is the S-cell used to extract features, corresponding to filtering operation of convolution kernel, one type is used to resist deformation C-cell, corresponding to activation function, pooling, etc.

2.2 Redesigned Mask Branch Structure

The upper-layer network structure of this paper is also divided into three parts. The first part is the training area recommendation network, the second part is the prediction category and the frame offset, and the third part is the redesigned mask branch network structure. Figure 1 is a redesigned block diagram of a mask branch network.

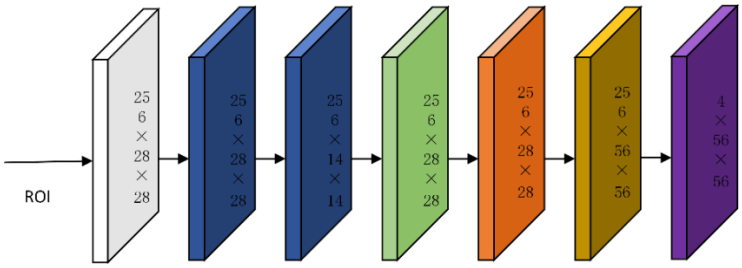

Figure 1. Mask Branch Convolution Structure

Scaled the input image to 1024×1024 size, extract feature by ResNet-101network, and obtain the (256, 28, 28) feature vector in RoIAlign layer by bilinear interpolation. This will retain more valid information, improve the accuracy of boundary segmentation. Used 256 3 ×3 convolution kernels, strides 2, get a feature map with a size of 14*14. There is no difference between using a normalized layer and not using it, so, this paper not use it, unified used Relu activation function. Green box is upsampling operation. Through the upsampling operation, the feature map in the second blue box is upsampled, and the resolution is expanded to 28×28, then use the 256 1×1 convolution kernels to extract features from the first blue box, and then add their corresponding pixels to the summation. The advantage of using a 1×1 convolution kernel is that can greatly reduce overfitting and helpful for training. The benefits of upsampling are enhance the combination of semantic information and shallow information, and improve the segmentation effect. Then used 256 3 ×3 convolution kernels, strides 1, further feature extraction. Yellow box is deconvolution operation to expand the feature map resolution, and the end is the mask for each category.

The loss function used to calculate the formula: L = Lcls + Lbox + Lmask. The class loss function Lcls and the bounding regression error Lbox are calculated in the same way as in the Faster R-CNN. Lmask is calculated as:

$$L_{\text{mask}}(Cls\_k) = Sigmoid\ (Cls\_k) \quad (1)$$

Cls_k is the mask corresponding to each category, this definition makes it unnecessary for the network to distinguish which class each pixel belongs to. It only needs to distinguish different subclasses in this category, effectively avoid competition between categories, calculate the corresponding mask loss for each category of interest region, use Sigmoid loss function for each pixel.

The modified model effectively improves the boundary segmentation. By using the combination of front and back layer features and increase the resolution of the RoIAlign layer, obtained accurate boundary information. Instance segmentation is for pixels, so, the feature resolution of the roialign layer is improved to 28×28, which preserves more pixels on the feature map than the previous feature resolution, which is beneficial for accurate boundary segmentation. The algorithm of this

paper has achieved good results.

3. Experiment

There are 3 types of experimental data sets, including pigs, cow, and sheep. There are 2000 training sets, 8,00 verification sets, and 434 test sets. We control the smallest side of each image to be greater than 300, and the largest side is less than 1024. The reason for this is because we want to scale the image to 1024×1024 as input. In addition, we create the corresponding mask label by using the labelme tool. Due to the small number of images, we used data enhancement technology to expand the data, flip the image, and the corresponding mask label will also flip. In addition, used pre-trained models previously trained on coco datasets. Data enhancement technology and pre-trained models can effectively prevent over-fitting problems.

Considering the limited GPU memory, we fixed the previous basic network part and only trained the upper network branch. When training the candidate box in the first stage, we select the candidate box with higher score, and then finally select 2000 candidate boxes by maximal value suppression. In the second stage, as with the parameters for training Faster R-CNN, select 64 from 2000 candidate boxes, and the positive and negative ratio is 1:3. If the overlapping area of the 2000 candidate frames and the target frame obtained by the first stage RPN network is not less than 0.5, it is a positive sample, otherwise it is a negative sample. In the third stage, the selected 64 candidate boxes are used for the mask branch, we only calculate the loss function on the candidate box of each positive sample. There are four types of data sets for this experiment.

In the test, the maximum detection value of 100 detection frames is obtained by maximal value suppression, and the mask branch is applied to the 100 detection frames. These detection frames can be parallelized by keras' TimeDistributed function, which greatly improves the operation speed. The mask output is then adjusted to the ROI size and binarized using a threshold of 0.5.

Used the keras framework in this experiment. The image is scaled to 1024×1024 when the image is input. The batch size is 1 image per GPU. After 100 iterations, the learning rate is 0.02, the weight attenuation is 0.0001, momentum is 0.9. Experimenting on the Nvidia GeForce GTX 1080 GPU took 8 days to train.

We use the redesigned model for instance segmentation and compare it with Mask-R-CNN. Figure 2 is a comparison of the instance segmentation obtained by Mask-R-CNN and the method of this paper.

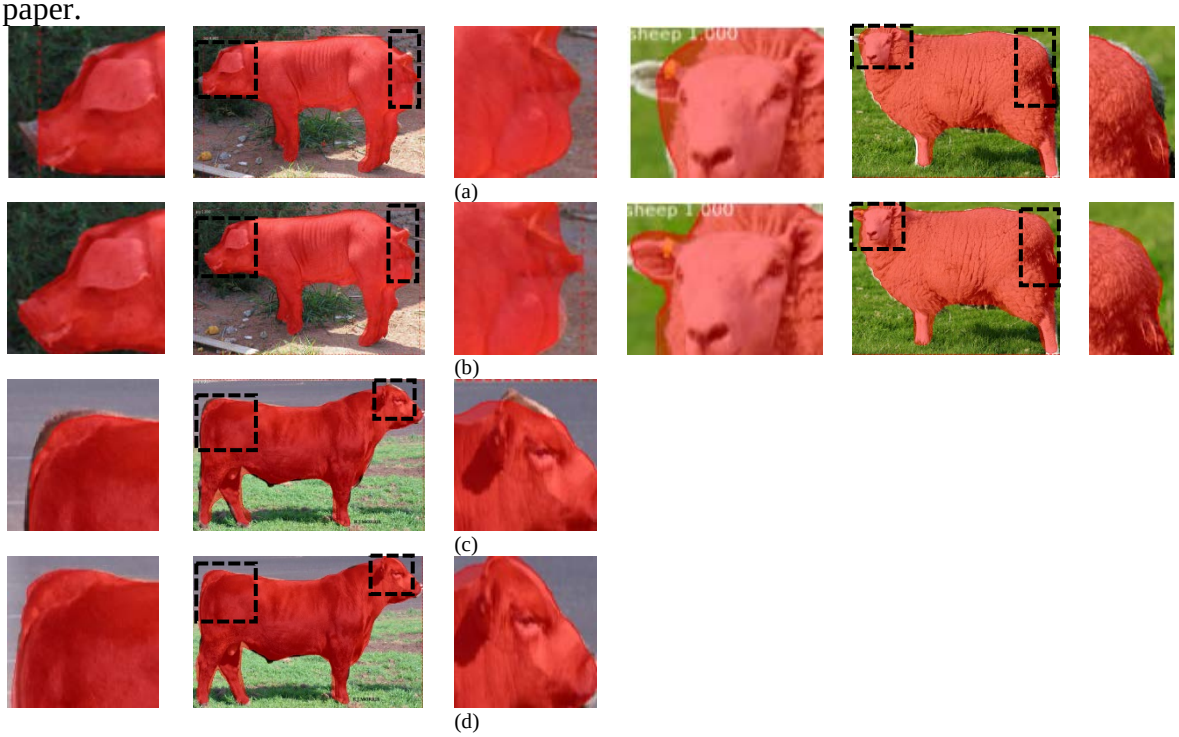

Figure 2. Comparison of Mask-R-CNN and the method of this paper.

(a),(c)Mask-R-CNN segmentation effect figure (b),(d) this paper segmentation effect figure

By comparing the segmentation map, the model in this paper can obtain more accurate boundary information, this is because the feature resolution of the RoIAlign layer is improved, which is increased to 28×28 by bilinear interpolation, and more pixel points on the feature map are retained. Used feature fusion on the mask branch to enhance the combination of semantic information and shallow information, and improve the segmentation effect.

We evaluated the results of this experiment through mAP. We calculate the IOU for each mask. The IOU is calculated as: IOU=(A∩B)/(A∪B). A is the mask corresponding to a candidate box, and B is the given mask label. Sort the IOU of each candidate box and mask label from largest to lowest, and set an IOU threshold. We set the threshold to 0.8, as long as there is one greater than this threshold and the corresponding category is the same that means the matching success, otherwise it is a match failure. According to the above matching results, the accuracy rate and the recall rate are respectively calculated, and the mAP is calculated according to the accuracy rate and the recall rate. We compare it with Mask-R-CNN. The results in Table 1 show that the method has higher precision.

By comparing the values of mAP in Table 1, the redesigned model is slightly better than Mask-R-CNN. Mask-R-CNN is a simple and effective instance segmentation system. It only adds a mask branch to the object detection and achieves a good segmentation effect. This paper redesigned the branch structure. Through the method of feature fusion and increased the resolution of the RoIAlign layer, improved the boundary segmentation. Table 1 shows the model that we designed has achieved a good segmentation effect.

TABLE 1. Average Accuracy Mean Comparison

Algorithm	**mAP（IOU=0.9）**
Mask-R-CNN	0.603
Ours	0.727

4. Conclusion

Based on the basic architecture of Mask-R-CNN, this paper redesigned the mask branch network structure. By using the combination of front and back layer features and increase the resolution of the RoIAlign layer, improved the boundary segmentation. Although we have improved the overall border segmentation, but it reduces the efficiency of network operation. This paper expanded the resolution of feature maps and used feature fusion, increased the amount of calculation. How to improve the efficiency of the algorithm while ensuring the accuracy of the algorithm, this is the focus of our future research.

References

[1] Hariharan B, Arbeláez P, Girshick R, et al. Hypercolumns for object segmentation and fine-grained localization[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2015:447-456.

[2] Dai J, He K, Sun J. Instance-Aware Semantic Segmentation via Multi-task Network Cascades[C]// Computer Vision and Pattern Recognition. IEEE, 2016:3150-3158.

[3] He K, Gkioxari G, Dollár P, et al. Mask R-CNN[C]// IEEE International Conference on Computer Vision. IEEE, 2017:2980-2988.

[4] Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]// International Conference on Neural Information Processing Systems. MIT Press, 2015:91-99.

[5] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[6] Lin T Y, Dollar P, Girshick R, et al. Feature Pyramid Networks for Object Detection[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2017:936-944.

[7] Hosang J, Benenson R, Dollár P, et al. What Makes for Effective Detection Proposals?[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2016, 38(4):814.

[8] Kirillov A, Levinkov E, Andres B, et al. Instancecut: from edges to instances with multicut[C]//CVPR. 2017, 3: 9.

[9] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2015:3431-3440.