Multi scale switchable atrous convolution for target detection based on feature pyramid

Cheng Fang*, *Ziqiang* Hao and *Jiaxin* Chen

School of Electronic Information Engineering, Changchun University of Science and Technology, Changchun, China

Abstract. Repeated observation mechanism can effectively solve the problem of low efficiency of feature extraction. By extracting features for many times to strengthen target features, this paper proposed a multi-scale switchable atrous convolution based on feature pyramid, SPC. The head of the detector adopted pyramid convolution mode, constructs 3-D convolution in the feature pyramid, and detected the same target in different pyramid levels by using the shared convolution with different stride changes, which realized the repeated observation of target features on multi-scale. The module optimized the convolution layer, extracted the features of the same image by convolution check of different sizes, and then selected and integrated the extracted results by using switch function, which effectively expanded the field of view of convolution kernel. In this paper, we choosed retinanet as the baseline network, and improved the loss function of focal loss proposed by retinanet to further solved the problem of unbalanced number of samples and sample distribution in the network model. The proposed method performed well on MS coco data set, improved the average accuracy of 9.8% on the basis of retinanet to 48.9%, and achieved FPS of 5.1 in 1333 * 800 images.

Keywords: Machine vision, Atrous convolution, Feature pyramid, Focal loss.

1 Introduction

Adding multiple viewing and thinking mechanism to machine learning can effectively enhance the detection performance. For example, fast r-cnn and its variants, the representative of the two-stage detector, first outputs the target proposal, then gives the target weight according to the loss function, and finally outputs the target detection result[1][2][3]. The disadvantage is that the detection speed is too slow. The goal of single-stage detection is to achieve the same detection accuracy as the two-stage detector with faster detection speed, and predict the position of the image intensively through scale and aspect ratio. Yolov3 extremely pursues the balance between speed and detection accuracy, which leads to the fact that both of them are not good enough[4]; SSD cancels the

* Corresponding author: 1748443494@qq.com

resampling of the pixels or features of the bounding box, which further improves the detection speed, but there is no special treatment for the target size in the feature extraction process;CenterNet chooses to set the detection point in the key position, and does not require NMS processing to repeat the preset box, which improves the detection speed, but the detection of key points will lead to target omission in the detection process[5].

To solve the above problems, this paper proposes a multi-scale switchable hole convolution based on feature pyramid. The module can be divided into two parts. The first part is multi-scale feature fusion on feature pyramid, and the second part is feature extraction of different scale targets using switchable hole convolution. The overall network takes advantage of the strong correlation between adjacent feature maps to extract features by sharing convolution kernels at all levels, so as to enhance the connection between scales and improve efficiency. At the same time, a switchable convolution is proposed to expand the convolution kernel field of view, so that different proportions of objects can be accurately recognized. According to the position information of the feature map, the feature map is fused, and the feature map generated by different levels of pyramid is matched by the step adjustment in space to complete the target detection. Then the improved loss function is introduced, and two parameters are added to improve the training effect. The experimental results show that all the improvements bring positive benefits to the network.

This paper adopts the single-stage detection Retinanet network, because the head of the RetinaNet network is more consistent with the design idea of the SPC model,and has stronger compatibility backbone network chooses the deeper ResNet-101 for experimentation. The loss function is also improved in response to the problems in the simple training, and the speed of the loss regression of the network model is accelerated.

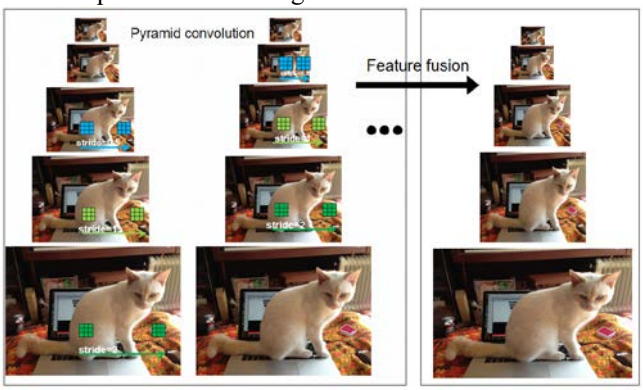

Fig. 1. Pyramid convolution as 3D convolution.

2 Multi scale switchable pyramid convolution

2.1. Pyramid convolution

Pyramid convolution (Pconv) is a spatial 3-D convolution composed of multi-scale 2-D convolutions.Pconv is equivalent to n different convolution kernels. Figure 1 describes the scale of feature pyramids at different levels, in which the image size of each level is only used for display scale, and does not represent the actual feature map. Three kinds of convolution kernels are used to construct 3-D convolution. When the pyramid level rises, the size of the feature image changes. The convolution kernel sampling with the same step will lead to the object mismatch. Therefore, when convolution is carried out in different layers, K asynchronous steps need to be set to ensure the same proportion.

For Pconv with n = 3, the first kernel step is equal to 2. If the third is equal to 0.5, the output is:

$$y^l = w_1 *_{s0.5} x^{l+1} + w_0 * x^l + w_{-1} *_{s2} x^{l-1} \tag{1}$$

where l is pyramid level, w1, w0 and w-1 are independent 2D convolution kernels, X is input feature map, and *s0.5 is convolution with 0.5 step. The kernel with 0.5 step is further replaced by the normal, the continuous convolution with 1 step and the continuous bilinear up sampling layer:

$$y^l = Upsample(w_1 * x^{l+1}) + w_0 * x^l + w_{-1} *_{s2} x^{l-1} \tag{2}$$

The above expression is fully expressed in the second layer. For the first level pyramid, the last term is set to zero. Similarly for the last level, the last term in the equation is set to zero. Because of repeated operations between adjacent levels, even if each level performs convolution calculation for three times, the actual amount of calculation is still only 1.5 times of the traditional method.

2.2. Multi scale switchable pyramid convolution

The proposed SPC convolutes the same input features with different atrous rates. The results are selected by the switch function, and then the features are fused. The selection function is determined by the size of the target in the space. The existing feature fusion methods reduce or enlarge the feature maps of different levels to the same resolution for simple summary, but completely ignore the differences of detail information and semantic information between high-level features and low-level features, and the influence of the obtained feature maps on the detection results is limited. In this paper, SPC is used to perform explicit convolution on the scale dimension to capture the interaction between scales, share the convolution kernel, and connect the single convolution kernel to achieve 3-D convolution on the image, so that the convolution kernel will expand with the different scale.

Firstly, the retinanet detector is analyzed, and it is found that the same object is not matched on the spatial scale in the detection process. Therefore, the convolution kernel size of the detector is changed to 3*3, and the proposed SPC structure is used to improve the detection ability. In order to distinguish the two functions, two additional convolution layers are added after the shared SPC module to meet the function of the detector.

The convolution kernel stride in the detector head is also changed from 1 to K asynchronous strides which can adapt to pyramid convolution structure. Because using the same training weight can not give full play to the feature extraction ability of switchable convolution in space, a convolution form with different training weights is proposed to train hole convolutions with different ratios. The expression of switchable convolution is as follows:

$$Conv(x,w,1) = S(x)\cdot Conv(x,w,1) + (1-S(x))\cdot Conv(x,w+\Delta w,r) \tag{3}$$

r is the void convolution rate, which is equivalent to introducing $r-1$ zeros between convolution kernels: $k_e = k + (k-1)(r-1)$. We choose $Conv(x,w,r)$ to represent the convolution kernel of 3*3, where w shows different forms under different conditions of R, so we need to ensure that the training weights of the two are different. When the void ratio is $r=1$, the convolution kernel is equivalent to the original convolution kernel. Experiments show that

the model performs best when $r=3$. Therefore, the convolution kernel weight of $r=1$ is set as w, and the convolution kernel weight of $r=3$ is expressed as $w+\Delta w$.

The switch function is used to select the objects in the image, and different sizes of objects are distinguished. The convolution kernel of corresponding scale and average pooling are used to extract the best features. In the process of network design, it is found that only using 7*7 convolution kernel can not give full play to the advantages of void convolution feature extraction, so we choose to design a pooling structure that matches the void ratio. Figure. 3 is a schematic diagram of switchable hole convolution (SAC).

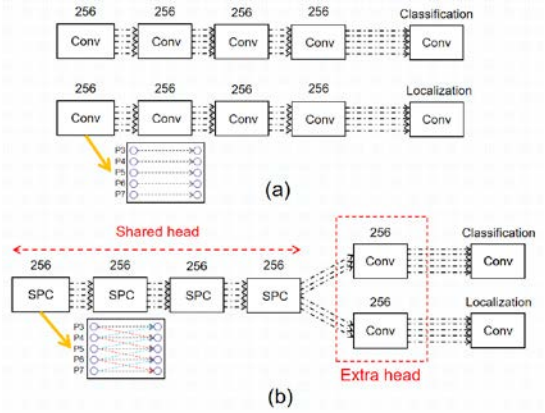

Fig. 2. (a) Structure of retinanet head. (b) Improved retinanet head.

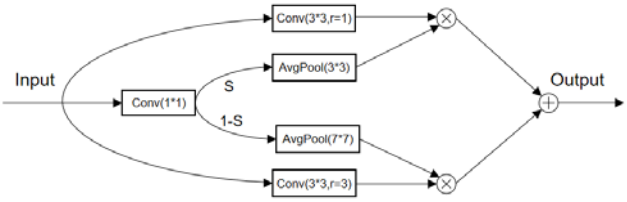

Fig. 3. Switchable atrous convolution (SAC).

The 3*3 convolution core in the detector is replaced by the switchable hole convolution, and the 3-D convolution is constructed by using the feature pyramid in space to realize the SPC network. When the super parameter n in the feature pyramid convolution is 3, the layer feature pyramid output of the SPC network is as follows:

$$\begin{aligned} y^l = {} & S(x^{l+1})Conv_1(x,w_1,1)_{s0.5}x^{l+1} + S(x)Conv_0(x,w_0,1)_{s1}x^l + S(x^{l-1})Conv_{-1}(x,w_1,1)_{s2}x^{l-1} \\ & + (1-S(x^{l+1}))Conv_1(x,w_1+\Delta w_1,1)_{s0.5}x^{l+1} + (1-S(x))Conv_0(x,w_0+\Delta w_0,1)_{s1}x^l \\ & + (1-S(x^{l-1}))Conv_{-1}(x,w_{-1}+\Delta w_{-1},1)_{s2}x^{l-1} \end{aligned} \tag{4}$$

3 Improved focal loss function

3.1. Sample distribution analysis

There are two common imbalanced distribution problems in the target detection dataset: (1) the number of samples of different categories is quite different; (2) the number of simple samples and difficult samples is quite different, which will have a significant impact on the

training and detection performance of the algorithm. Problem (1) will cause the network to each kind of training result has the obvious superiority and inferiority disparity. The situation of problem (2) is more obvious, simple samples will lead to the simplification of target detection performance, and the detection performance of the trained model for difficult targets is greatly reduced.

Focal loss introduced in Retinanet model was improved in this paper. By setting different loss function weights for candidate boxes with different confidence degrees, the detection effect on difficult samples was improved. The experiment proved that the algorithm improved the detection performance for difficult positive samples.

3.2. Loss function design based on focal loss

In the training process, the imbalance between foreground and background is one of the main reasons that affect the detection accuracy of target detection algorithm. In this paper, we improve the loss function of focal loss to solve the problem of sample imbalance.

The traditional cross entropy function formula is as follows:

$$CE(p,y)=\begin{cases}-\log p & if\ y=1\\ -\log(1-p) & otherwise\end{cases} \tag{5}$$

where y is the sample label, $y=1$ is the positive sample, $y=-1$ is the negative sample, and p represents the probability that the network predicts that the sample is a positive sample. In order to simplify the binary classification cross entropy loss function, P_t is defined as follows:

$$P_t=\begin{cases}p & if\ y=1\\ 1-p & otherwise\end{cases} \tag{6}$$

In this case, the cross entropy loss function can be abbreviated as:

$$CE(p,y)=CE(p_t)=-\log(p_t) \tag{7}$$

In order to solve the problem of positive and negative sample imbalance in sample imbalance, focal loss assigns a weight factor α to the positive and negative samples,The value range of α is 0~1, and α is set to be inversely proportional to the number of positive and negative samples to reduce the impact of the imbalance of positive and negative samples on network performance. By adding a super paramete α to the cross entropy loss function, we can solve the class imbalance problem of the training data set. For the problem of the imbalance of the difficult and easy samples, we can add a weight factor $(1-P_t)^{\gamma}$ to solve the problem of the difficult and easy samples. The final focal loss expression is as follows:

$$FL(P_t)=-\alpha_t(1-P_t)^{\gamma}\log(P_t) \tag{8}$$

Based on the idea of focal loss, the following loss function is designed:

$$
\begin{aligned}
loss = {} & \lambda_{coord} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} [(x_i + \hat{x}_i^{\ j})^2 + (y_i + \hat{y}_i^{\ j})^2] + \\
& \lambda_{coor} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} [(\sqrt{w_i^j} - \sqrt{\hat{w}_i^j})^2 + (\sqrt{h_i^j} - \sqrt{\hat{h}_i^j})^2] + \\
& \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} [\hat{C}_i^j \log(C_i^j) + (1 - \hat{C}_i^j) \log(1 - C_i^j)] + \\
& \lambda_{\ noobj} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{noobj} [\hat{C}_i^j \log(C_i^j) + (1 - \hat{C}_i^j) \log(1 - C_i^j)] + \\
& \sum_{i=0}^{S^2} I_{ij}^{obj} \sum_{c \in classes} [\hat{P}_i^j (1 - P_i^j)^{\gamma} \log(P_i^j) + (1 - \hat{P}_i^j)(p_i)^{\gamma} \log(1 - P_i^j)]
\end{aligned} \tag{9}
$$

The loss function consists of five parts: (1) the sum of squares of the predicted target center and the actual label box center (2) The square sum of the length and width of the prediction box and the actual target is lost (3) The entropy loss is the cross entropy loss of the category containing the target (4) It is the loss of cross entropy for the class without target (5) Cross entropy loss of confidence.Parameter I_{ij}^{obj} indicates whether the j-th candidate box of the i-th grid contains the prediction of the target, If there is a target $I_{ij}^{obj} = 1$, otherwise it is 0. Similarly, the parameter I_{ij}^{noobj} indicates whether the candidate box does not contain the prediction of the target, $I_{ij}^{noobj} = 1 - I_{ij}^{obj}$.If the center of the object falls into each grid, the grid is said to contain the target. The parameters x, y, w and h represent the center coordinates (x, y) of the prediction frame and the length and width (w, h) of the target frame respectively. The values are normalized and range from 0 to 1.Parameter C represents the category number of the prediction, and parameter p represents the confidence level of the prediction. By setting λ_{coor} and λ_{noobj} two parameters to balance the imbalance between foreground category and background category, This article is set as follows: $\lambda_{coor} = 5$ and $\lambda_{noobj} = 0.5$.By adding γ value to the cross entropy loss part to increase and weaken the dominant role of more simple samples on the loss function, the detection effect of the network for difficult samples is better. Experiments show that the general γ value of 2 is better. The experimental results verify the effectiveness of the improved loss function.

4 Experiment

4.1. Experimental details

The experiment was conducted on the coco data set, the training set was MS coco train2017, and the number of pictures in the training set was about 118K. Using mmdetection to realize retinanet, the network model is SPC, the backbone network is resnet-101, choose to use a single GPU NVIDIA geforce RTX 3090 for training, and set 4 pictures for GPU training, batch size is 128. In the training process, 12 epochs were selected, the initial learning rate was set to 0.01, and multiplied by 0.1 at the 8th and 11th epochs. The long side of the input image is adjusted to 1333, and the short side is adjusted to 800, without changing the aspect ratio of the image.

In the evaluation of other function fusion modules, the parameters used are consistent with the training parameters of SPC network in this paper, and only the feature fusion part of other networks is replaced, and no additional modules are used. The difference of

detection accuracy between feature fusion modules is more obvious. The comparison results are shown in table 1.

4.2. Experimental results and analysis

Firstly, the SPC module is compared with the current mainstream feature fusion method, and the test results are shown in Table 1. It can be seen that the SPC module has the highest detection accuracy, which is improved by 3.8% compared with Libra, and the amount of data calculation is reduced by 22%. Compared with PA net network, the average accuracy is improved by 4.3%. The results show that the SPC feature fusion method is effective to improve the retinanet head, and the feature extraction and fusion on multi-scale can effectively improve the detection accuracy of the detector.

Table 1. The comparison between SPC and other function fusion modules is carried out on coco minival.

Feature fusion	AP	AP50	AP75	FLOPS(G)
FPN	38.5	57.3	41.2	239.3
HR-Net	38.6	57.1	41.3	297.6
PA-Net	38.9	57.6	41.6	245.9
NAS-FPN	39.1	57.0	41.8	347.1
Libra	39.4	58.7	42.2	315.8
SPC	43.2	61.3	46.0	244.7

Figure 4 shows the visualization results of the network model. The SPC module can accurately identify the real object, and the size of the detection box is the best. It is also obvious in the results that due to the multi-scale feature detection, the advantage of the large field of view of SPC is that the detection effect of large target is more obvious, It can ensure that the global context information has a positive impact on the detection of occluded objects in the process of feature transfer.

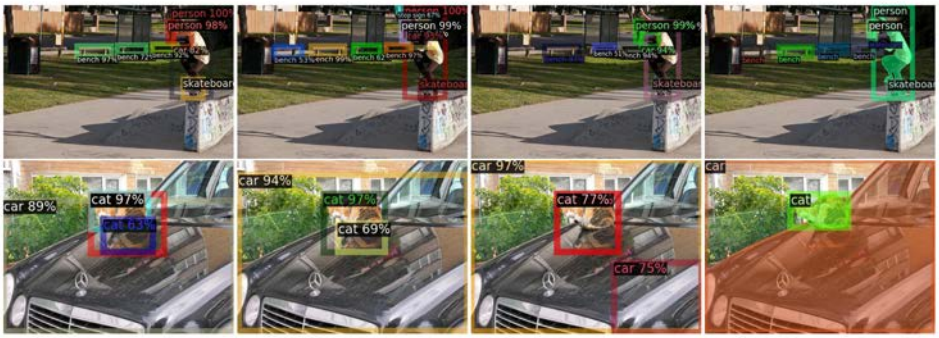

Fig. 4. From left to right: visual detection results through FPN, Pconv, SAC and SPC.

Figure 5 shows the RetinaNet network with all parts of the proposed module losses during the period of training, the results show that the improved loss function improved in the process of network training samples of the imbalance problem of the impact of, training a good example for weight reducing at the same time increase the difficulty of the training effect is poorer sample weight, effectively reduces the loss of network training.

The overall network test is carried out on coco test dev, including the proposed SPC module and the improved loss function. At the same time, in order to achieve the highest detection accuracy, we also add some additional modules, such as the common regularization method dropblock, integrated BN connection. Table 2 shows the comparison

results between this network and other mainstream networks. It can be seen from the results that the improved loss function improves the network performance by 0.4-0.5AP. Compared with the benchmark model retinanet, SPC network improves 9.8 AP and 3.2 AP, reaching the highest performance of 48.9 AP. FPS is 5.1, which is a significant improvement, It is proved that good results have been achieved in feature fusion.

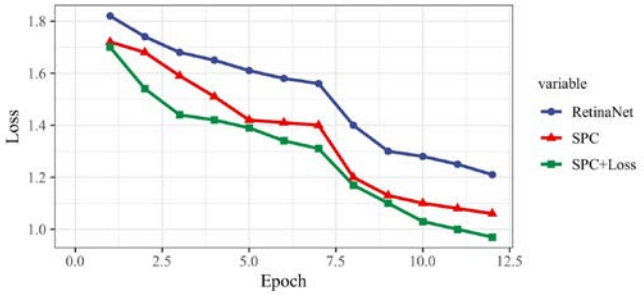

Fig. 5. Training loss of each part in 12 training periods.

Table 2. Latest comparison of coco test dev for bounding box object detection.

* Represents the use of an improved loss function							
Method	Backbone	AP	AP50	AP75	APS	APM	APL
Yolov3	DarkNet-53	33.0	57.9	34.4	18.3	25.4	47.9
RetinaNet	ResNet-101	39.1	59.1	42.3	21.8	42.7	50.2
Libra R-CNN	ResNet-101	42.6	62.9	45.8	23.7	44.2	54.1
PA-Net	ResNet-101	45.7	65.3	48.6	28.2	49.3	47.6
Ours	ResNet-101	48.4	69.1	52.3	32.3	52.1	63.6
Ours*	ResNet-101	48.9	69.5	52.7	32.7	52.4	64.1

5 Conclusion

SPC uses the feature correlation between adjacent levels to extract features from multi-scale, and fuses features through switch function, so that the model can automatically select the appropriate feature map according to the size of the location target. The model can adapt to the feature pyramid that the feature map becomes smaller as the level increases, Using asynchronous amplitude to align the scale change of feature map in space, 3-D convolution is realized in scale and space dimension, which greatly improves the efficiency of target detection. The improved loss function can make the model converge rapidly, and set different weights for different degree of difficulty samples to increase the accuracy of target location. Compared with the baseline network, the whole network improves 9.8ap, and the detection speed also meets the expectation of single-stage detector, reaching 5.1fps. The SPC network designed in this paper has achieved competitive results in target detection.

References

1. Kaiming He 2015.Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in Neural Information Processing Systems, pages 91–99

2. Zhaowei 2018 Proc. Int. Conf. on Computer Vision and Pattern Recognition, passage 6154-6162
3. Jiangmiao Pang 2019 Libra r-cnn :Towards balance learning for object detection.Proc. Int. Conf. On Computer Vision and Pattern Recognition, passage 821-830
4. Joseph Redmon and Ali Farhadi 2018 Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767
5. Xingyi Zhou and Philipp Krahenbuhl 2019 Objects as points.arXiv preprint arXiv:1904.07850,2019