www.czasopisma.pan.pl PAN www.journals.pan.pl
INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2019, VOL. 65, NO. 2, PP. 293–301
Manuscript received March 3, 2019; revised April, 2019. DOI: 10.24425/ijet.2019.126313

Detecting objects using Rolling Convolution and Recurrent Neural Network

WenQing Huang, MingZhu Huang, and YaMing Wang

***Abstract*—At present, most of the existing target detection algorithms use the method of region proposal to search for the target in the image. The most effective regional proposal method usually requires thousands of target prediction areas to achieve high recall rate.This lowers the detection efficiency. Even though recent region proposal network approach have yielded good results by using hundreds of proposals, it still faces the challenge when applied to small objects and precise locations. This is mainly because these approaches use coarse feature. Therefore, we propose a new method for extracting more efficient global features and multi-scale features to provide target detection performance. Given that feature maps under continuous convolution lose the resolution required to detect small objects when obtaining deeper semantic information; hence, we use rolling convolution (RC) to maintain the high resolution of low-level feature maps to explore objects in greater detail, even if there is no structure dedicated to combining the features of multiple convolutional layers. Furthermore, we use a recurrent neural network of multiple gated recurrent units (GRUs) at the top of the convolutional layer to highlight useful global context locations for assisting in the detection of objects. Through experiments in the benchmark data set, our proposed method achieved 78.2% mAP in PASCAL VOC 2007 and 72.3% mAP in PASCAL VOC 2012 dataset. It has been verified through many experiments that this method has reached a more advanced level of detection.**

***Keywords*—multi-scale features, global context information, rolling convolution, recurrent neural network**

I. Introduction

With the development of society, the application of target detection tasks in social life is increasingly in dispensable. In the early days, people tried to use template matching to detect objects. However, due to the large difference in the target category and the complex background, the performance of such methods is not ideal. There is very little work in this area. At present, the research results of target detection methods based on deep learning are very advantageous. In the traditional object detection algorithm, C.Wohler and J. Anlauf combined with the feed forward neural network and local receptive fields (LRF) to construct a more successful target detection system [1]. In 2005, N. Dalal and B. Triggls proposed the famous directional gradient histogram HOG feature descriptor [2]. A number of improved studies were subsequently proposed for their results. Literatures[3], [4], [5] used integral maps to calculate HOG to enhance HOG performance. Literature [6] uses a multi-scale deformable combined model as a detector to achieve better target detection by using a more powerful detector. There are also hybrid features and combined detectors to improve target detection performance, such as the literature [7] combined optical flow characteristics and HOG for pedestrian detection in the video, the literature [8] combined Adaboost and SVM to form a two-stage detector. However, the traditional target detection algorithm sliding window method selects the window strategy is not targeted, basically adopts an exhaustive way, takes a long time, and has a lot of redundancy. In addition, the characteristics of manual design can not meet the target detection requirements of diversity, which leads to the bottleneck of the traditional target detection algorithm, and the actual need of target detection is not achieved in speed and accuracy. In recent years, the deep convolutional neural network (CNN) model [9], [10], [11], [12], [13] has shown amazing advantages in many research fields, and the application of target detection technology has made this technology level a big step forward. In the target detection algorithm based on convolutional neural network, the successful application of convolutional neural network (R-CNN) [14] based on regional proposal in the field of target detection has quickly triggered the research boom of target detection technology. In a class of target detection methods[15], [16], [17] based on regional proposals, different search methods are used to select a large number of target prediction regions in the image, and then feature extraction is performed on these target prediction regions using CNN.Based on the above research basis, two subsequent improved network models, fast R-CNN [18] and fast R-CNN [19], have once again made great progress in detection efficiency. These two network models share similar pipelines, and achieve synchronization positioning and classification by feeding the features obtained by CNN to the region of interest (ROI) pool layer and then simulating the multitasking loss function. Recent research [20], [21], [22] demonstrates the validity of this view.

A. Multi-scale feature fusion

At present, most prior art target detection methods use a regional nomination method to search for objects in an image. The most effective regional proposal method usually requires a large number of target prediction areas to effectively search

This work was supported by the Natural Science Foundation of Zhejiang Province (LZ15F020004),the Natural Science Foundation of National(61272311) and 521 Project of Zhejiang Sci-Tech University.

WenQing Huang is with School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou, China; (e-mali: patternrecog@163.com)

MingZhu Huang is with School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou, China; (e-mali: 851489278@qq.com)

YaMing Wang is the correspondence author and is with School of Information Science and Technology, Zhejiang Sci-Tech University, Hangzhou, China; (e-mali:yamingwang2000@163.com)

for the correct target location. Although recent research has been able to achieve good detection results by generating hundreds of target prediction regions, it still faces challenges in the application of small target detection. This is mainly due to the fact that such algorithms only utilize top-level feature maps with high semantics but low resolution. With the continuous development of target detection technology, many researchers have used multi-layer feature maps to obtain multi-scale local context information to improve detection performance when applying convolutional networks for target detection. For example, the R-FCN [23] algorithm calculates semantic segmentation by accumulating partial scores for multiple regions of interest for each category. Hypercolumes [24] uses a similar approach for object segmentation. Several other methods (HyperNet [25], ParseNet [26], and ION [27]) select target prediction box that incorporate multiple layers of feature information for classification and location. This is equivalent to the summation conversion function. SSD [28] and MS-CNN [29] are also based on multi-level feature map prediction target boxes. Through these studies, it is not difficult to find that the high-resolution feature information of the low-level convolution is favorable for searching small target positions, and the high-semantic feature information of the high-level convolution can be well used for target classification. Therefore, by combining the feature information of the upper layer and the lower layer, the detection performance of the small target can be effectively improved. However, these methods all adopt convolution structure with special fusion multi-scale features, which makes the network structure more complex, and the time consumption will also become longer because of the special fusion processing process. Therefore, this paper proposes a rolling convolution method that combines the semantic information of higher convolution layers with the feature information of lower convolution layers in a continuous convolution process. And no special multi-scale feature convolution structure is added to make the network structure simple and functional.

B. Global context information

In addition to multi-scale features, the study also shows that global context information also has a certain auxiliary effect on the correct classification of targets. Global information is a general description of the entire image. The appearance of the target in the global scene information has a close dependence, and different scenarios always imply the possibility of different targets appearing. So the global context provides a rich clue for target detection, providing important information in the target classification process. It is suitable for a variety of visual recognition tasks. At present, in the target detection algorithm of convolutional neural networks, there are usually two ways to add global context information: One is to average the pool of the entire feature map to obtain global context features; and the other is to use the Recurrent Neural Network (RNN) to process the correlation between the sequence information to select favorable context information for the target region. The global draw pooling will add the average value of each cell of the entire feature map as a global feature. Such a process is too computationally simple, and does not take into account that there must be a lot of noise interference in the entire feature map. These noises are not beneficial to the detection process, so that obtaining global features has limited help for detection. The premise of the RRN design network model is based on the inter-relationship between elements. Therefore, many studies have introduced loop neural networks into target detection to extract effective context information for target detection. This kind of network has ”memory”. Unlike traditional neural networks, RNN can use ”sequence information” to analyze the correlation between elements, so it is used to select favorable context information for the target. The existing representative algorithms are mainly ION[24] algorithm and AC-CNN[29] algorithm. The ION algorithm divides the feature map and uses the RNN independently from the top, bottom, left, and right directions. After the information propagation in the four directions is performed separately, the information in the four directions needs to be merged (1*1conv). The AC-CNN algorithm is another convolutional neural network based on the attention context mechanism. It combines traditional fast CNN with contextual features and fuses multi-scale local and global features into a single network framework. In order to obtain global information, the attention-based context sub-network generates favorable position-dependent feature maps by using the stacked LSTM layers as target areas. Through their research, it has been shown that the target detection effect based on the cyclic neural network to extract the global context is much better than that based on the global flat pool to obtain the global information. Based on AC-CNN research and analysis, it shows that LSTM-based cyclic neural network has great advantages in extracting global context. The LSTM-based RNN effectively solves the fact that the RNN cannot handle remote dependencies, and the GRU is a variant of the LSTM. The GRU maintains the characteristics of the LSTM and is simpler in terms of structural design. Therefore, this paper introduces a cyclic neural network based on GRUs to extract more effective global context information for the network.

C. summarize

On this basis, a multi-feature research direction combining with global features is proposed. Based on the Faster R-CNN network model, this paper uses the multi-scale feature extraction method of rolling convolution to extract global features using GRNN-based RNN. The entire training process is carried out in an end-to-end manner, and we will verify the effectiveness of the method by comparing the Pascalvoc benchmark dataset. The main research results of this paper can be summarized as follows: 1)A new target detection network structure combining multi-scale local features and global context features is proposed. 2) We obtain multi-scale local features using rolling convolution. Compared to applying a special structure that extracts multi-layer features, our method is more concise and detection is more stable. 3) We propose a RRN model to extract useful information on the global feature map to assist in detection; thereby, avoiding the limitation of global average pooling.

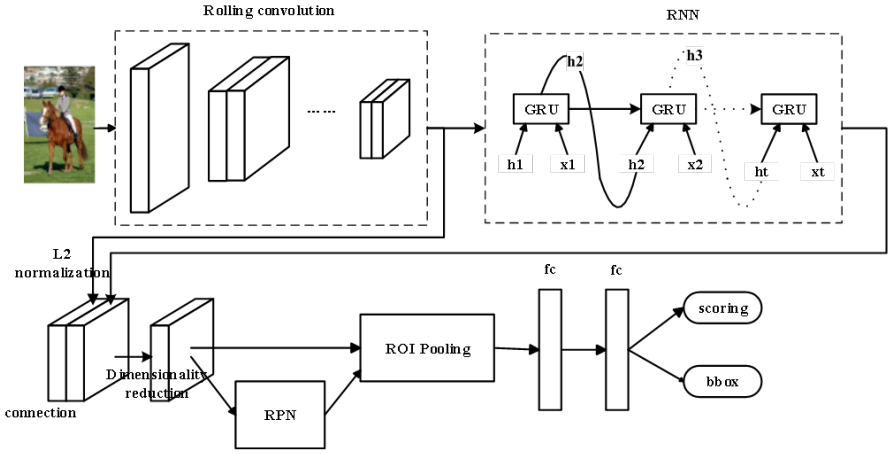


Fig. 1: Application of context information for object detection. The rolling convolution structure convolution processing the input image to obtain multi-scale features. Then, a recurrent model with three GRU layers is used to repeatedly detect the useful area from the global view. After L2 normalization, cascading, scaling, and dimension reduction, the feature maps of the global context information and multi-scale information are sent to the RPN network selection region proposals. Each target prediction area is sent to two fully connected layers for classification and regression.

II. ARCHITECTURE

In this section, we describe a detector structure that applies multi-scale feature to the region proposal using rolling convolution. Furthermore, it introduces the valued global context information using an RNN to help improve detection performance. Firstly, we provide a brief summary of the entire network and then discuss the network in detail.

During the detection process, the original picture is input to the RC layer, and an activation map is generated by the convolution operation. The activation map deepens the semantic information and preserves the feature information of input layer (relatively higher resolution) using rolling convolution. The GRU is then added at the top of the convolution such that the global context information is gradually and selectively introduced into the context feature maps. Then, we combine the context feature map with the multi-scale feature map, and process the feature maps by L2 norm normalization and compress them into a unified space. The Regional Proposal Network (RPN) then generates approximately 100 target prediction areas based on the fused feature map. The Region of interest pooling (ROI Pooling) layer will pool the generated target prediction area to a fixed scale. Finally, the detection module classifies and locates these predicted regions.

A. *Multi-scale feature based on Rolling Convolution*

An effective target detection system should be able to detect correctly for multiple types and sizes of targets. In the region-based detector, the type setting of the target prediction boxes is limited, which inevitably leads to the inability to fully comply with all types of target detection requirements. At the same time, the resolution of the feature map of the output of the final convolutional layer is much smaller than the resolution of the input image. Therefore, the feature information of the small target becomes inconspicuous and thus difficult to detect. Thus, detecting small objects becomes a problem because the features that represent the fine details of small objects in low resolution feature maps may be weak. Improving detection using multi-scale features is one way to alleviate this problem. There are various methods for combining multi-scale features, such as inputting multi-scale images, or assembling multi-layer features to construct a feature map pyramid structure. Many published experiments have shown that this method improves detection performance, especially for small or overlapping objects. Through these experiments, we found that multi-scale features satisfy the following conditions: Sufficiently high resolution to represent a more detailed part of the object; high-level semantic information can be extracted by converting the input image into a feature map; context information facilitates object classification and boundary regression. Through these analyses, we improved the convolution method without adding additional structures to obtain multi-scale features, as shown in Figure 2.

We use the following expressions to explain how the rolling convolution process extracts multi-scale information:

$$Q_n = r_n\{\mu_n(Q_{n-1}), p_n(Q_{n-1})\} \tag{1}$$

$$size(p_n(Q_{n-1}) = size(Q_n) \tag{2}$$

Where Q_n is a feature map in the n layer, and un is a nonlinear operation for mapping features in the $(n-1)$ layer to features of the n layer. $p_n()$ is a pool function applied to the feature map. Its purpose is to match the size of the upper feature map to the next map. $r_n()$ is used to reduce the required computation by reducing the dimensions of the combined results of convolution and direct aggregation results. $size()$ represents the scale of a two-dimensional feature map. Thus, the convolution of each layer not only obtains deeper semantic information by nonlinear transformation, but it can also retain all the information of the feature map of the upper layer by direct pooling. The top-level feature maps

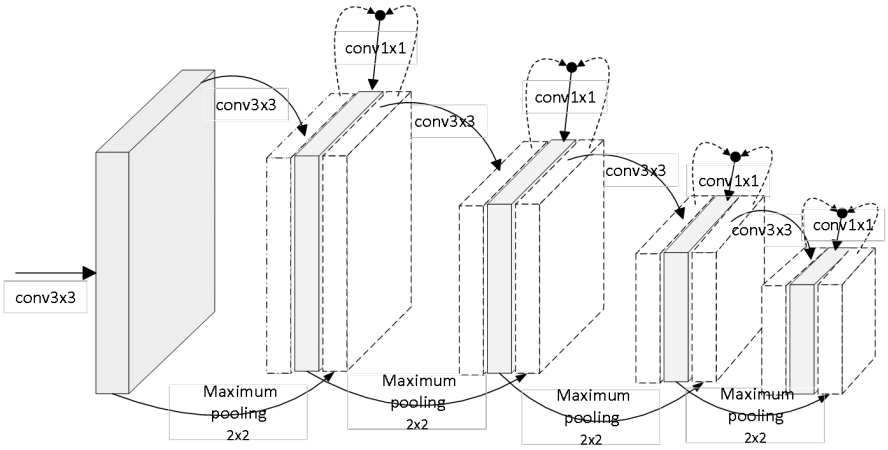


Fig. 2: The structure of rolling convolution. Each layer of convolution is completed in three parts. First, in order to obtain higher semantic information for target classification, a 33 convolution channel is used to generate the feature maps of the next layer, thereby deepening the semantic information of the feature; then, the largest pooling layer is on the upper layer. The output features are pooled to retain the feature information of the previous layer; finally, the 11 convolution is performed to reduce the dimension of the feature maps connected in the first two steps. In addition, all of the feature maps (solid boxes) in the figure represent the output features of each layer, and the direction indicated by the arrows is the direction of convolution or pooling, and is shown by the dashed box.

contain different levels of semantic information and matching resolution of all layers. Furthermore, it is not simply spliced together; instead, it is fused at every layer. Ultimately, the only requirement is that candidate box extracted from the top-level feature map must contain the multi-scale information of the corresponding location.

B. Global context information based on RNN

For object detection, viewing a picture from a global perspective can provide a lot of useful information. In order to explore useful global context information, we apply an RNN model to select valuable global context information to improve the detection of objects. This RNN model consists of three GRUs with the top-level feature map being used as an input. Feature learning through RNN models is used to highlight effective context enhancing detection performance. We will now explain how to build the RNN model to obtain global context information. The RNN model is composed of three layers of GRUs. We have applied the GRU implementation method discussed in [30]. Let $Q = (q_i, ..., q_t)$ be the input feature maps set, where $q_i(i = 1, ..., t)$ has the D dimensions. The proposed GRUs consists of four weight matrices W_l, W_r, G_l and G_r. At each recurrent level t, the activation function of the hidden unit $h_j(t)$ of the $j-th$ layer can be expressed by the following equation:

$$h_j^{(t)} = w_c \bar{h}_j^{(t)} + w_l \bar{h}_{j-1}^{(t-1)} + w_r \bar{h}_{j-2}^{(t-2)} \tag{3}$$

where w_c, w_l, and w_r are the values of a gate with a sum of 1. The hidden unit is initialized as follows:

$$h_j^{(0)} = U x_j \tag{4}$$

where U projects the input into a hidden space. The new activation is computed as follows:

$$\bar{h}_j^{(t)} = \phi(W^l h_{j-1}^t + W^r h_j^t) \tag{5}$$

where is an element-wise nonlinearity. The gating coefficients w_c, w_l, and w_r are computed as follows:

$$\begin{bmatrix} w_c \\ w_l \\ w_r \end{bmatrix} = \frac{1}{Z} exp(G^l \bar{h}_{j-1}^{(t)} + G^r \bar{h}_j^{(t)}) \tag{6}$$

where G^l, $G^r \in R^{3*d}$ and

$$Z = \sum_{k=1}^{3} [exp(G^l h_{j-1}^{(t)} + G^r h_j^{(t)})] \tag{7}$$

In each GRU layer, a position-dependent weight map $l_{i,j}$ is predicted according to the input feature sequence, that is, the Softmax value of kk positions in the input feature map. This is the probability estimate calculated by the RNN based on the correlation between the sequences. The larger the probability value, the corresponding region corresponding to the sequence is favorable for the target classification. The relevant weight map $l_{i,j}$ is calculated as follows, where w_j is the weighting operation on the $j-th$ element, and x_t is the input feature sequence, which can take any sequence in the kk position.

$$l_{i,j} = p(x_t = j | x_{t-1}, ..., x_1) = \frac{exp(w_j h^{(t)})}{\sum_{j=1}^{K^2} exp(w_j h^{(t)})} \quad j \in \{1, ..., K^2\} \tag{8}$$

Finally, the global context feature maps can be obtained by multiplying the obtained position-related weight maps and the input feature maps. where X_t is the feature maps and $X_{t,i}$ is the $i-th$ slice of the feature maps at time-step t.

TABLE I: Comparison of test results on PASCAL VOC 2007 and PASCAL VOC 2012

Methods	PASCAL VOC 2007				PASCAL VOC 2012			
	Fast R-CNN	Faster R-CNN	HyperNet	Ours	Fast R-CNN	Faster R-CNN	HyperNet	Ours
mAP	70.0	73.2	76.3	78.2	67.5	70.4	71.4	72.3
aero	77.0	76.5	77.4	80.3	75.3	84.9	84.2	88.0
bike	78.1	79.0	83.3	84.1	76.8	79.8	78.5	76.5
bird	69.3	70.9	75.4	78.5	70.8	74.3	73.6	74.9
boat	59.4	65.5	69.1	70.8	52.3	53.9	55.6	58.5
bottle	38.3	52.1	62.4	68.5	32.7	49.8	53.7	50
bus	81.6	83.1	83.1	88.0	77.8	77.5	78.7	76.7
car	78.6	84.7	87.4	85.9	71.6	75.9	79.8	83.2
cat	86.7	86.4	87.4	87.8	89.3	88.5	87.7	89.8
chair	42.8	52.0	57.1	60.3	44.2	45.6	49.6	50.1
cow	78.8	81.9	79.8	85.2	73.0	77.1	74.9	75.2
table	68.9	65.7	71.4	73.7	55.0	55.3	52.1	53.1
dog	84.7	84.8	85.1	87.2	83.5	86.9	86.0	88.5
horse	82.0	84.6	85.1	86.5	80.5	81.7	81.7	82.4
mbike	76.6	77.5	80.0	85.0	80.8	80.9	83.3	85.4
person	69.9	76.7	79.1	76.4	72.0	79.6	81.8	76.3
plant	31.8	38.8	51.2	48.5	35.1	40.1	48.6	48.5
sheep	70.1	73.6	79.1	76.3	68.3	72.6	73.5	72.1
sofa	74.8	73.9	75.7	75.5	65.7	60.9	59.4	63.2
train	80.4	83.0	80.9	85.0	80.4	81.2	79.9	85.7
tv	70.4	72.6	76.5	81.0	64.2	61.5	65.7	68.3

TABLE II: Comparison of two sets of test results on PASCAL VOC 2007

Methods	multi-scale features		Global feature	
	Super-feature	rolling convolution	global average pooling	RNN model
mAP	39.0	39.7	41.5	42.4
acro	51.2	49.1	55.0	53.7
bike	52.6	52.3	53.9	54.9
bird	32.7	33.1	31.2	33.0
boat	27.4	25.6	26.3	26.4
bottle	18.9	25.0	25.5	29.7
bus	55.9	53.1	60.5	66.1
car	51.7	54.2	52.0	53.6
cat	50.7	49.3	45.1	48.2
chair	21.4	25.6	20.6	26.7
cow	45.9	45.1	55.3	46.5
table	36.8	39.1	33.0	37.7
dog	43.9	46.8	56.1	47.1
horse	31.7	38.9	44.2	45.1
mbike	41.4	38.3	44.3	40.8
person	25.5	20.5	24.7	28.3
plant	15.4	19.5	11.8	18.2
sheep	36.2	35.8	35.8	37.9
sofa	43.0	41.0	39.7	44.4
train	44.1	49.8	45.5	50.6
tv	53.2	52.1	56.6	58.3

$$x_t = E_p(x_t = j|x_{t-1}, ..., x_1)[X_t] = \sum_{i=1}^{K^2} l_{t,i} X_{t,j} \qquad (9)$$

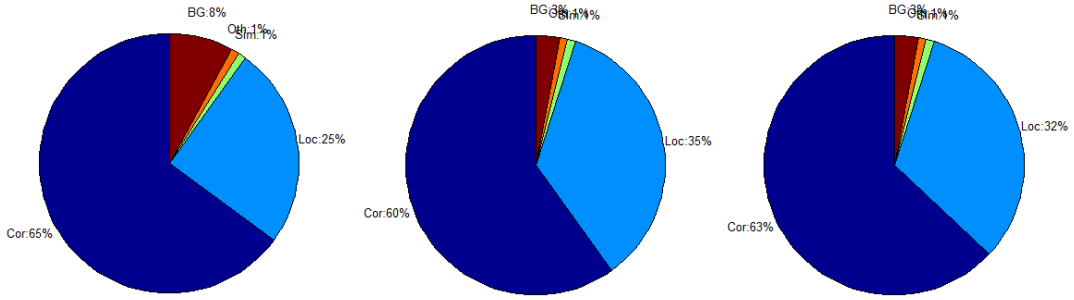
III. RESULTS

A. Experimental setup

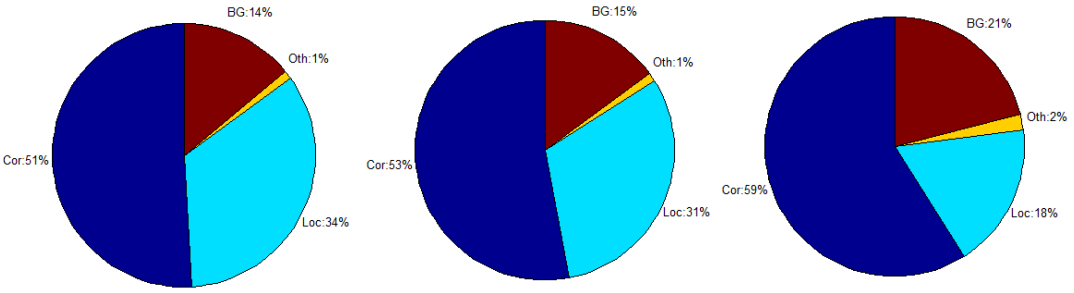
The method uses the Faster R-CNN network model on the Caffe platform and initializes the first five layers of convolution parameters of the network with pre-trained VGG16 [4] convolutional parameters. The parameters of all newly added layers were initialized by a zero mean Gaussian distribution with standard deviations of 0.01 and 0.001, respectively. During the training process, the random gradient descent (SGD) is used to train and adjust the network parameters. We use four images as a batch, generate 128 ROIs for each image over the RPN network, and generate 512 ROI sample training update parameters for each iteration. In each batch, 25% of the regional samples are set to foreground, and the rest are used as background training examples. And the Intersection over Union is (IOU) set to be greater than or equal to 0.5. In terms of data enhancement, this paper flips the image with a probability of 0.5 to enhance the training data. The following parameters were set in the calculation to update the network parameters: a maximum of 70000 iterations, a momentum of 0.9, a learning rate of 0.001, and a 10x reduction per 2000 iterations. On hardware devices, we train the Nvidia Geforce GTX 1070 GPU and 12GB of RAM.

B. Performance Comparisons

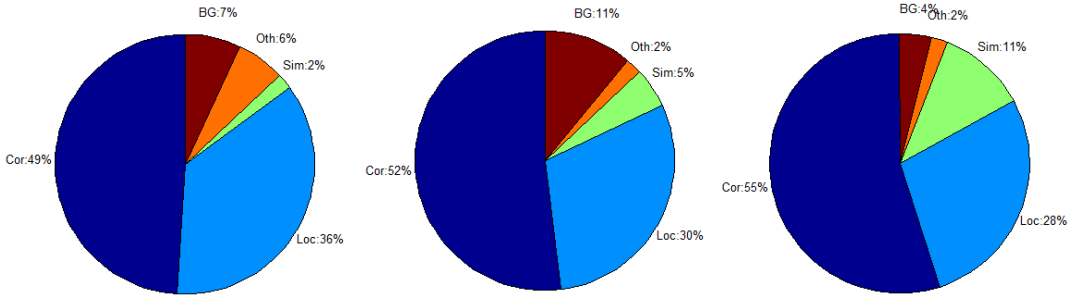
This article uses the benchmark datasets: PASCAL VOC 2007 and PASCAL VOC 2012 to verify the performance of the proposed network. These two data sets are common target



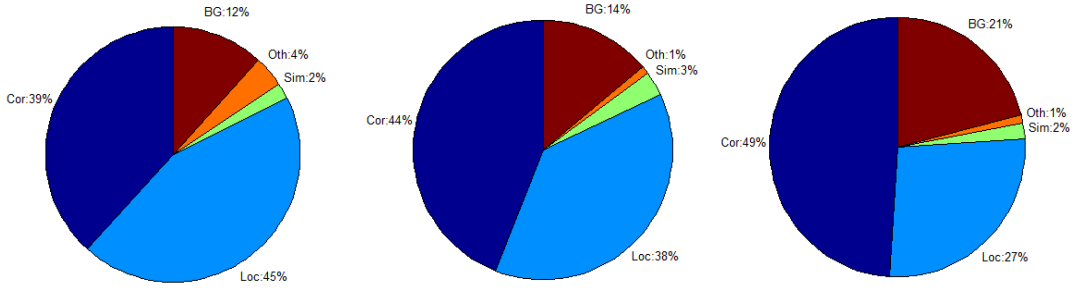
(a) boat



(b) bottle



(c) chair



(d) plant

Fig. 3: Statistical comparison of the four target detection results with the lowest accuracy rate on Pascal VOC. Among them, the pie chart is mainly based on the correct detection (Cor), positioning error (Loc), error caused by confusion with other targets (Sim), confusion with unmarked targets (BG) and other (Oth). In the figure, the pie charts from left to right represent the Fast R-CNN, the Faster R-CNN, and the statistical results of the methods in this paper.

Fig. 4: Effect of the three methods on the same picture. Fast R-CNN uses a cyan detection frame, and Faster R-CNN uses a light blue detection frame. The method proposed in this paper uses a red detection frame.

detection annotation data sets with 20 target classes. There are 9,963 images in the PASCAL VOC 2007 dataset and 11,540 images in the PASCAL VOC 2012.

We will propose a detection algorithm compared with Fast R-CNN, Faster R-CNN and HyperNet on the two data sets respectively, with Mean Average Precision (mAP) as a metric. Table 1 shows the results of the proposed method, Fast R-CNN, Faster R-CNN, and HyperNet on two benchmark data sets, respectively. On the Pascal VOC 2012, the Fast R-CNN algorithm based on selective search achieves 70% mAP, the Faster R-CNN algorithm achieves 73.2% mAP, and the HyperNet algorithm achieves 76.3% mAP. The algorithm in this paper achieved 78.2% mAP, which improved 8% mAP, 5% mAP and 2% mAP, respectively, compared with fast R-CNN, faster R-CNN and Hypernet. The experimental results show that the feature map extracted by this method is more accurate than the feature map extracted by other methods. The test results on the Pascal VOC 2012 dataset show that compared with the Fast R-CNN, the Faster R-CNN and the HyperNet, the detection algorithm proposed in this paper increases the by 5% mAP, 2%mAP and 1% mAP respectively. Based on the above data, we can draw a similar conclusion: the method of this paper has good detection performance in target detection.

C. Detailed Analysis

In order to further analyze our method, two sets of comparative experiments were performed on Pascal VOC 2007, and the IOU value was set to be greater than or equal to 0.7. In the network structure of this paper, we use the method of rolling convolution to obtain multi-scale features. To test whether this method is useful, we compare this method with a super-feature method based on joining multiple layers of features. The data in Table 2 shows that multi-scale features that rely on rolling convolution are 0.7% mAP higher than detection of super features that are connected to multi-layer features. The method based on rolling convolution has no superiority in the detection results, but is structurally simpler. At the same time, in order to test whether the method in this paper is effective in global feature extraction, we compare it with the global average pooling method. From the results, the global information obtained by the method in this paper is superior to the global average pooling method in detection performance. The reason is that the global average pooling method may introduce some interference information, thus reduce the detection accuracy. The use of a RNN model can be used to optimize the associated position to highlight the positive impact of these areas on target detection. Therefore, we use GRUs to achieve global context calculations more reasonable. Figure 3 shows the statistical comparison of the four target detection results with the lowest accuracy rate on Pascal VOC. In the figure, the pie charts from left to right represent the Fast R-CNN, the Faster R-CNN, and the statistical results of the methods in this paper. As can be seen from the statistical results, our method can significantly reduce the false positives of the categories with detection challenges. For example, in the two challenging categories of bottles and potted plants, they usually belong to the category of small targets. The detection algorithm proposed in this paper increases mAP by 5% and 10%, respectively. To a certain extent, this reflects the advantage of the feature acquisition method in this paper in small target detection.

IV. CONCLUSIONS

This paper proposes a new target detection network. Multi-scale features are extracted by using a rolling convolution method, and global context features are constructed using the GRUs-based RNN model. At the same time, we take the two parts of the acquired features into one multi-feature. The feature of the target region based on multi-feature selection will have a strong feature base, which can improve the detection performance of the target. In this paper, several experiments have been carried out in the dataset Pascal VOC, and the results show that our method does have an effect in detecting performance improvement.

REFERENCES

[1] WoHler C, Anlauf J K. An adaptable time-delay neural-network algorithm for image sequence analysis[J]. IEEE Transactions on Neural Networks, 1999, 10(6):1531-1536.
[2] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]// Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. IEEE, 2005:886-893.
[3] Laptev I. Improvements of Object Detection Using Boosted Histograms[C]// British Machine Vision Conference 2006, Edinburgh, Uk, September. DBLP, 2006:949-958.
[4] Shet V D, Neumann J, Ramesh V, et al. Bilattice-based Logical Reasoning for Human Detection[C]// Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on. IEEE, 2007:1-8.
[5] Zhang L, Wu B, Nevatia R. Detection and Tracking of Multiple Humans with Extensive Pose Articulation[C]// IEEE, International Conference on Computer Vision. IEEE, 2007:1-8.
[6] Azizpour H, Laptev I. Object Detection Using Strongly-Supervised Deformable Part Models[M]// Computer Vision ECCV 2012. Springer Berlin Heidelberg, 2012:836-849.
[7] Dalal N, Triggs B, Schmid C. Human detection using oriented histograms of flow and appearance[J]. 2006, 3952:428-441.
[8] Dollar P, Wojek C, Schiele B, et al. Pedestrian Detection: An Evaluation of the State of the Art[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(4):743.
[9] Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: Neural Information Processing Systems. (2012)11061114
[10] Lin, M., Chen, Q., Yan, S.: Network in network. In: International Conference on Learning Representations. (2014)
[11] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D.,Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. arXiv preprintarXiv:1409.4842 (2014)
[12] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scaleimage recognition. arXiv preprint arXiv:1409.1556 (2014)
[13] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition.arXiv preprint arXiv:1512.03385 (2015)
[14] Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. (2014) 580587
[15] Uijlings, J., van de Sande, K., Gevers, T., Smeulders, A.: Selective search for object recognition. International Journal on Computer Vision 104(2) (2013) 154171
[16] Zitnick, C.L., Doll ar, P.: Edge boxes: Locating object proposals from edges. In:European Conference on Computer Vision. (2014) 391405
[17] Arbelez, P., Pont-Tuset, J., Barron, J., Marques, F., Malik, J.: Multiscale combinatorial grouping. In: IEEE Conference on Computer Vision and Pattern Recognition. (2014) 328335
[18] Girshick, R.: Fast r-cnn. In: IEEE Conference on Computer Vision and Pattern Recognition. (2015) 14401448
[19] Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Neural Information Processing Systems.(2015) 9199

[20] Liang, X., Wei, Y., Shen, X., Jie, Z., Feng, J., Lin, L., Yan, S.: Reversible recursiveinstance-level object segmentation. arXiv preprint arXiv:1511.04517 (2015)
[21] Zeng, X., Ouyang, W., Wang, X.: Window-object relationship guided representation learning for generic object detections. arXiv preprint arXiv:1512.02736 (2015)
[22] Gidaris, S., Komodakis, N.: Object detection via a multi-region and semanticsegmentation-aware cnn model. In: IEEE International Conference on ComputerVision. (2015) 11341142
[23] Long, J., Shelhamer, E., Darrell, T. Fully convolutionalnetworks for semantic segmentation. In CVPR, 2015.
[24] Hariharan, B., Arbelez, P., Girshick, R., Malik, J. Hypercolumns for object segmentation and fine-grained localization. In CVPR, 2015.
[25] Kong, T., Yao, A., Chen, Y., Sun, F. Hypernet: Towards accurate region proposal generation and joint object detection. In CVPR, 2016.
[26] Liu, W., Rabinovich, A., Berg, A.C. ParseNet: Lookingwider to see better. In ICLR workshop, 2016.
[27] Bell, S., Zitnick, C.L., Bala, K., Girshick, R. Inside-outside net: Detecting objects in context with skip poolingand recurrent neural networks. In CVPR, 2016.
[28] Cai, Z., Fan, Q., Feris, R.S., Vasconcelos, N. A unified multi-scale deep convolutional neural network for fast object detection. In ECCV, 2016.
[29] Li J, Wei Y, Liang X, et al. Attentive Contexts for Object Detection[J]. IEEE Transactions on Multimedia, 2017, 19(5):944-954.
[30] Stewart, R., Andriluka, M. End-to-end people detection in crowded scenes. arXiv preprint arXiv:1506.04878 (2015).