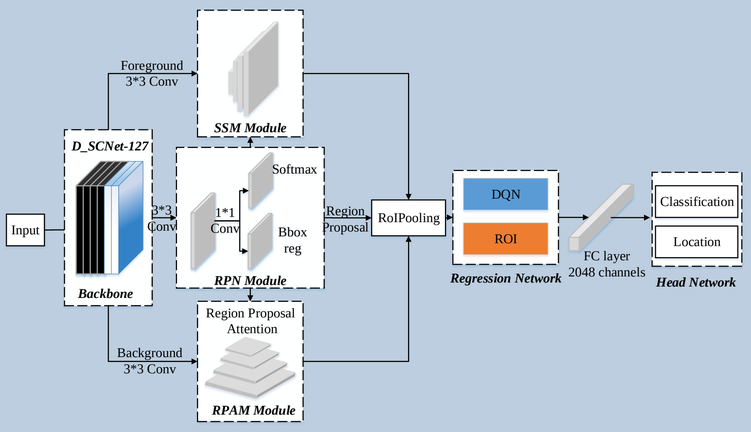
Object Detection Model Based on Scene-Level Region Proposal Self-Attention

Yu Quan, Zhixin Li*, Canlong Zhang, Huifang Ma
*Corresponding Author. E-mail: lizx@gxnu.edu.cn

Abstract

In order to improve the performance of two-stage object detection and consider the importance of scene and semantic information for visual recognition, the neural network of object detection algorithm is studied and analyzed in this paper. The main research work of this paper includes:

- We propose a deep separable convolution network named SCNet-127 R-CNN.
- We build the scene-level region proposal self-attention module.
- We propose a bounding box regression network module.



Visualization

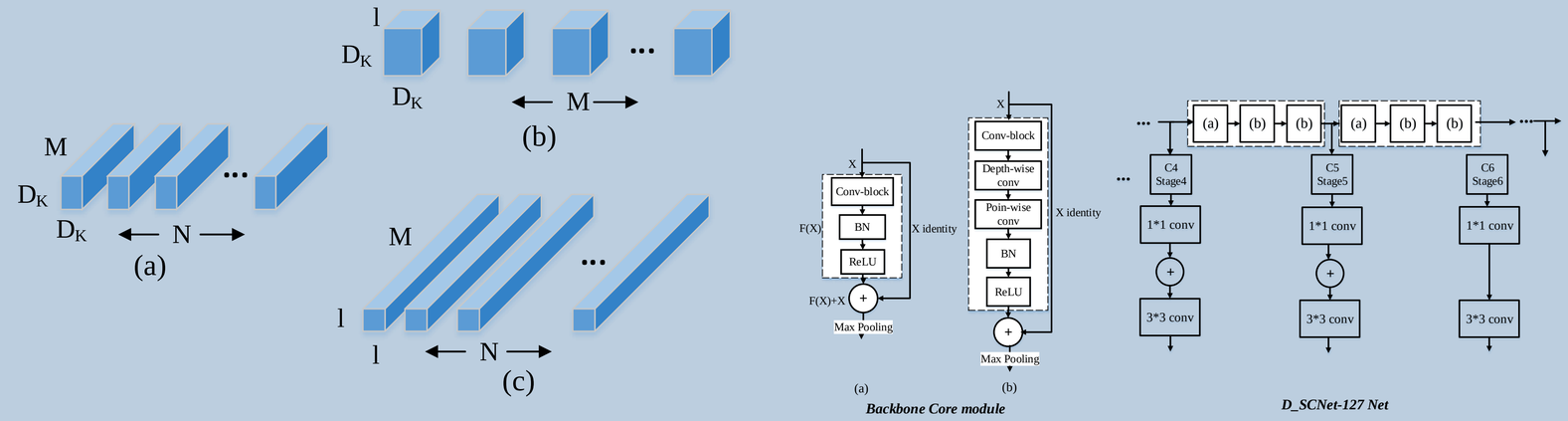
Object detection and positioning against a simple background. The green box represents the candidate area generated by DQN network each time, the red box represents the final positioning result obtained by combining the regression networks, and the white box represents the real target area.

Acknowledgements

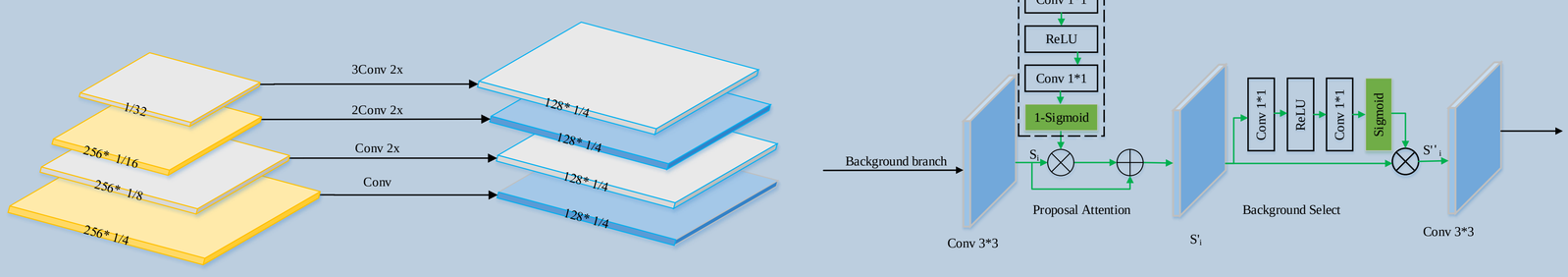
This work is supported by the National Natural Science Foundation of China (Nos. 61966004, 61663004, 61762078, 61866004), the Guangxi Natural Science Foundation (Nos. 2019GXNSFDA245018, 2018GXNSFDA281009, 2017GXNSFAA198365), Guangxi "Bagui Scholar" Teams for Innovation and Research Project, Guangxi Collaborative Innovation Center of Multi-Source Information Integration and Intelligent Processing.

Proposed Model

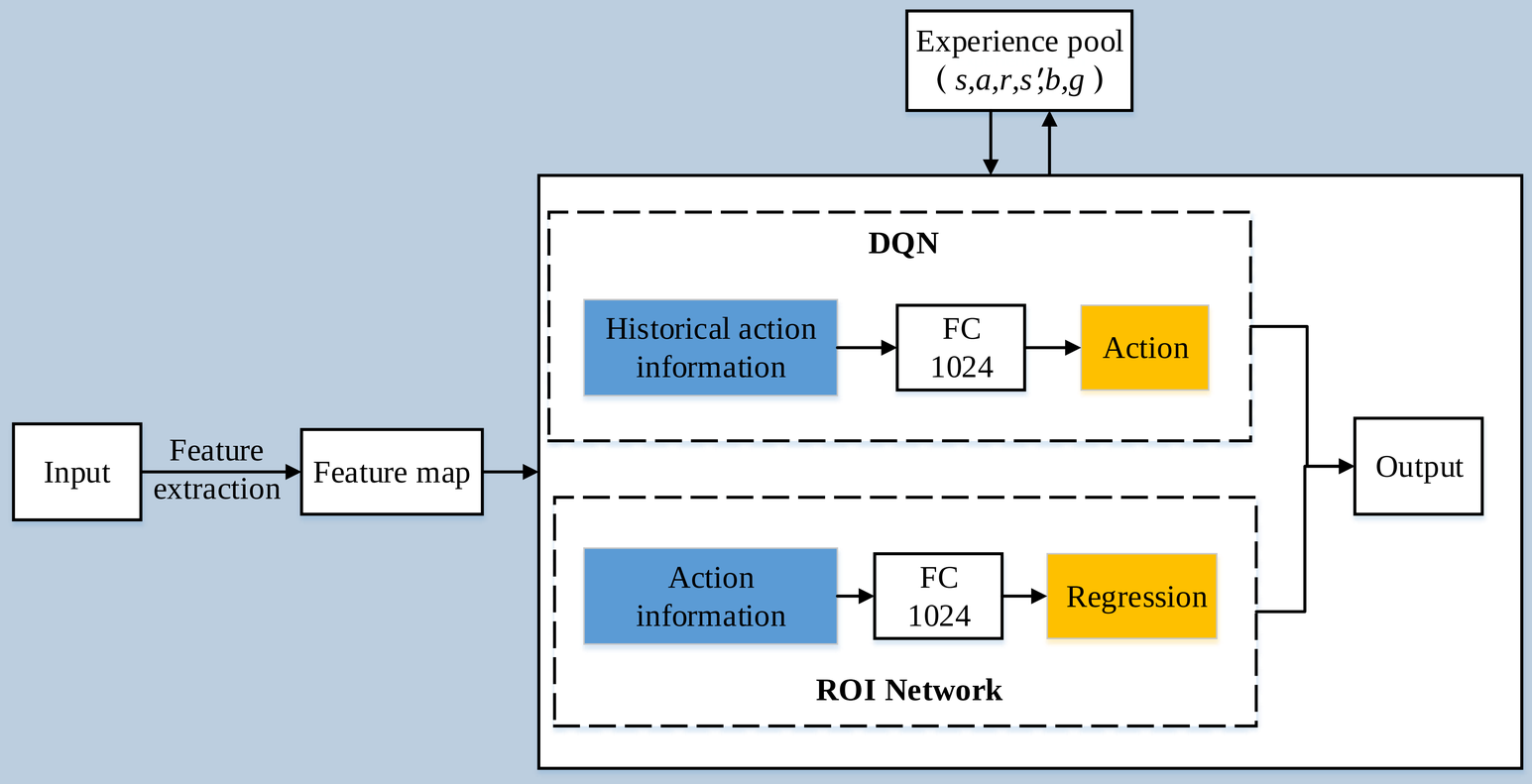
In order to reduce the time complexity of the model, we repeat the convolutional block by separable convolutional network unit as shown in below. (a) is a standard convolutional layer filter, and (b) and (c) is a depth convolution and a 1×1 convolution of a depth separable convolution filter, respectively. In order to achieve an accurate and fast multi-scale, multi-category image object detection behavior and obtain the accurate location information and category information of the object from the input image, the network model is reconstructed based on the process of region proposal. The right figure shows the structure of the backbone network based on the depth separable convolution.



For the details about Scene-Level Region Proposal Self-Attention Module, the SSM branch can obtain stronger semantic features, improve the performance of object detection, all levels of information from the FPN are combined into a single output to achieve high-density prediction. The RPAM branch introduces self-attention mechanism. The self-attention module combines useful information from the RPN branch and makes the detection task focus more on the local object to promote the accuracy of background semantics.



For the details about Bounding Box Regression Network Module, we show it as



Experiments

Several pre-trained basic ImageNet models and our own network models were used in the experiment. Among them, the VGG16 network model is called V16. The ResNet-50 network model is called R50. The table also shows the training time (Train Speed), the training rate (Test Rate), the test speed (Test Speedup), and the VOC07 and MSCOCO17 datasets. The average accuracies of models are compared.from the table, we can see that our methods are superior to other baseline methods.

TABLE II
RESULTS OF IMPACT OF VARIOUS BACKBONE NETWORKS ON FPN GPU-BASED TRAINING AND TEST RATE ANALYSIS RESULTS FOR MULTIPLE MODELS MSCOCO (%).

Evaluation	R-CNN		Fast R-CNN		Faster R-CNN		Mask R-CNN		D_SCNet-127 R-CNN		SSD		$YOLO_{V3}$	
	V_{16}	R_{50}	V_{16}	R_{50}	V_{16}	R_{50}	V_{16}	R_{50}	V_{16}	R_{50}	V_{16}	R_{50}	V_{16}	R_{50}
Train Time(h)	84	75	9.5	8.0	8.7	7.7	44	15	12.5	11.3	10.0	8.4	10.4	6.4
Train Speedu(×)	1	1.12	8.8	10.5	9.6	10.9	1.9	5.6	8.7	9.6	8.4	10.0	8.0	13.0
Test Rate(s/im)	47	5	0.32	0.25	0.14	0.11	0.2	0.07	0.047	0.03	0.045	0.038	0.047	0.029
Test Speedup(×)	1	9.4	146	188	335	427	522	671	1025	1375	1044	1236	1000	1620
Voc07+COCO17	45.7	49.6	65.1	67.5	70.4	77.6	70.5	77.3	81.6	84.5	76.3	78.1	57.9	73.8