Efficient-Receptive Field Block with Group Spatial Attention Mechanism for Object Detection

Jiacheng Zhang, Zhicheng Zhao, Fei Su

Beijing University of Posts and Telecommunications

Beijing Key Laboratory of Network System and Network Culture, Beijing, China

INTRODUCTION

北京郵電大學
Beijing University of Posts and Telecommunications

Feature Enrichment Module

Multibranch Structure
(different branches often correspond to different receptive fields)

When dilated rate and object size are mismatched, the discontinuity of convolution area will lead to inferior feature extraction.

Efficient Receptive Field Block (E-RFB) provides a sufficient RF by downsampling and increasing depth.

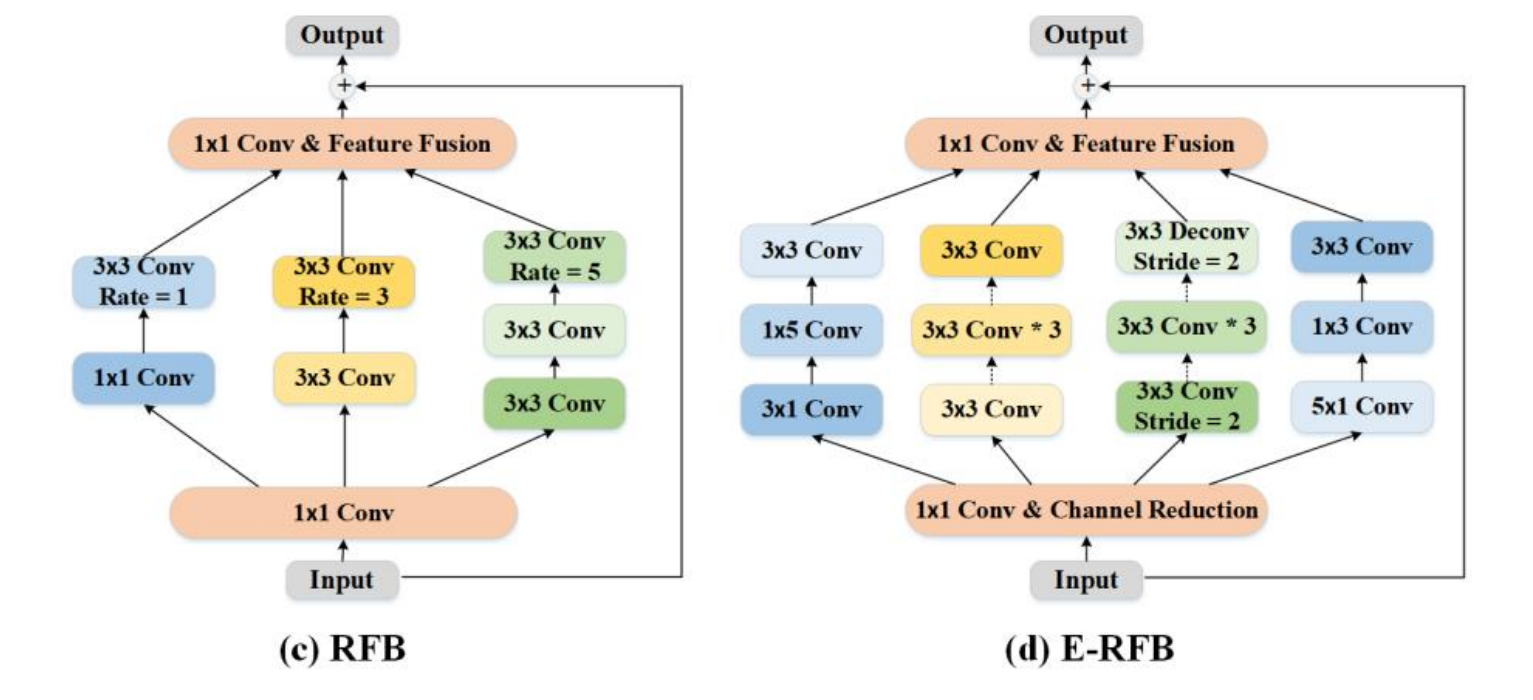
Attention Mechanism

The inconsistencies across different branches exist. The attention mechanism can filter conflictive information.

Pooling along a single dimension numerically have not consider the internal relationships of a feature map.

A spatial attention mechanism (GSAM) is proposed to gradually “slim” the attention feature map via channel grouping.

METHOD (Efficient-Receptive Field Block)



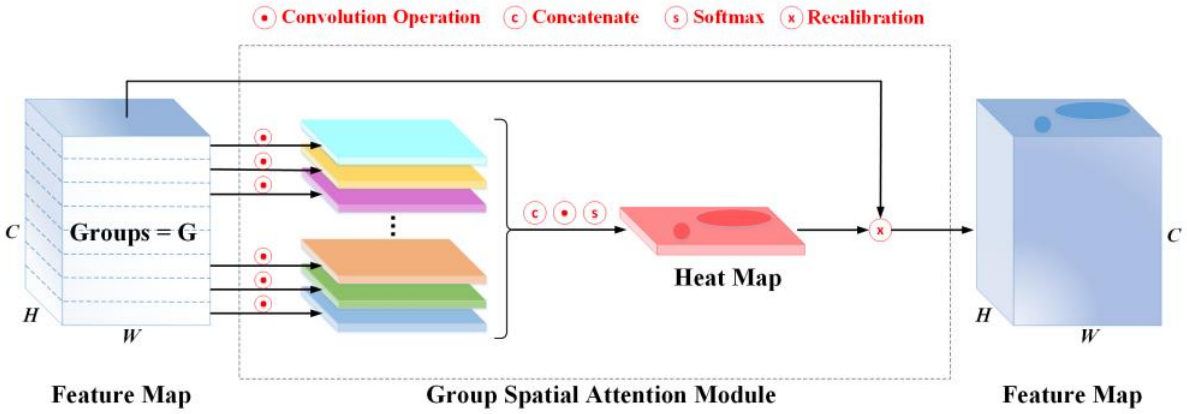
RFs similar to RFB.

Deep, narrow, and large cardinality.

Strip convolution.

- The ERFB is a five-branch structure with shortcut. Multiscale contexts are aggregated.
- A 1 × 1 convolution with a channel reduction of 16 is employed in each branch to reduce channels and form a bottleneck structure.
- Multiple cascaded convolutional layers are used to obtain a large RF.
- The n × m convolution is implemented by combining an n×1 and a 1×m conv layer.

METHOD (Group Spatial Attention Module)



(a) Structure of GSAM

(b) Visualization of heat maps

- Concatenated 3 × 3 group convolutions are used to obtain a three-dimensional attention map with G channels
- A 3 × 3 convolution is used to reduce dimension, and a softmax function is used as the gate function.
- The two-dimensional spatial attention heat map is obtained, which is used to recalibrate the input features.
- (b) shows the visualized heat maps, where a brighter pixel corresponds to a higher activation value.
- We show the heat map of the first GSAM output, which has been resized for better visualization.

The ratio of additional parameter when the GSAM is added to ordinary k × k convolution layer is $r_s \approx 1/c_2$.
The additional FLOP ratio can be calculated as $r_s \approx 1/c_2 + 1/2 \cdot k^2 \cdot c_1$, where c1 and c2 denote the number of input and output feature map channels.

EXPERIMENTS

北京郵電大學 Beijing University of Posts and Telecommunications

Experiments on PASCAL VOC and Microsoft COCO

TABLE I
QUANTITATIVE COMPARISON OF DETECTION METHODS ON THE TEST SET OF PASCAL VOC 2007.

Backbone	Model	Param.	mAP
MobileNet	SSD [22]	11.46M	70.47
	RFB [10]	6.75M	70.61
	Ours (E-RFB)	**6.27M**	71.31
	Ours (E-RFB + GSAM)	6.34M	**71.83**
VGG-16	SSD [22]	42.41M	78.93
	RFB [10]	36.53M	79.84
	Ours (E-RFB)	**28.74M**	79.95
	Ours (E-RFB + GSAM)	28.77M	**80.23**
ResNet-50	SSD [22]	37.23M	78.71
	RFB [10]	32.44M	79.45
	Ours (E-RFB)	**25.06M**	79.74
	Ours (E-RFB + GSAM)	25.09M	**80.61**

TABLE II
DETECTION PERFORMANCE ON THE COCO 2017 TEST-DEV DATASET.

Backbone	Model	Param.	FLOPs	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AR_S	AR_M	AR_L
MobileNet	SSD [22]	12.38M	2.21G	20.0	35.5	20.2	1.9	19.6	36.2	4.0	30.7	52.1
	RFB [10]	7.68M	**1.57G**	20.8	36.4	21.1	1.8	20.1	37.3	4.5	33.0	54.8
	Ours (E-RFB)	**6.91M**	1.69G	21.3	37.3	21.7	2.3	21.9	37.4	5.1	33.7	53.6
	Ours (E-RFB + GSAM)	6.99M	1.76G	**22.0**	**38.0**	**22.4**	**2.4**	**22.1**	**38.6**	**5.3**	**34.3**	**55.3**
VGG-16	SSD [22]	50.98M	42.65G	29.2	48.2	30.8	11.2	31.5	44.5	16.3	44.7	59.2
	RFB [10]	45.10M	39.92G	30.3	49.4	32.0	11.8	31.9	46.7	17.3	45.9	62.3
	Ours (E-RFB)	**35.57M**	**36.97G**	30.9	49.7	32.7	12.5	33.1	46.7	17.7	46.3	62.0
	Ours (E-RFB + GSAM)	35.59M	36.98G	**31.8**	**50.7**	**33.8**	**12.9**	**33.7**	**47.9**	**18.7**	**47.1**	**63.0**
ResNet-50	SSD [22]	45.81M	42.26G	32.3	51.7	34.6	14.7	35.6	45.6	21.2	50.0	62.1
	RFB [10]	41.0M1	40.12G	33.6	53.3	35.9	15.6	37.1	48.5	22.5	51.5	**64.8**
	Ours (E-RFB)	**31.77M**	**36.32G**	34.1	53.8	36.7	15.9	37.7	48.1	22.6	52.1	64.0
	Ours (E-RFB + GSAM)	31.80M	36.33G	**34.4**	**54.2**	**36.9**	**16.1**	**38.0**	**48.7**	**22.9**	**52.3**	64.3

The vanilla convolution layer in the detection head is replaced with the E-RFB module. The integration of E-RFB and GSAM is simple, that is, the latter uses the E-RFB's output feature map as input to model the heat map, and then uses the calculated heat map to recalibrate the features.

- Our detectors surpassing the baseline SSD by a large margin regardless of the backbone.
- Compared with RFB Net, E-RFB Net has higher accuracy with fewer parameters.
- Benefiting from its large RF, GSAM captures context information in a large range, making for large object detection.

EXPERIMENTS

Ablation Study

TABLE III
COMPARISON OF OTHER STATE-OF-THE-ART MULTIBRANCH BLOCKS ON VOC 2007 TEST SET AND COCO 2017 VAL SET. THE BACKBONE IS VGG-16.

Dataset	PASCAL VOC			MS COCO				
Model	Param.	FLOPs	mAP	Param.	FLOPs	AP	AP_{50}	AP_{75}
Residual [1]	**27.63M**	**32.73G**	78.74	**35.20M**	**36.52G**	24.4	42.6	24.8
ASPP(v) [9]	29.80M	33.78G	79.53	38.37M	37.57G	24.1	42.5	24.3
Inception [40]	29.18M	33.35G	78.86	37.76M	37.14G	24.3	42.0	24.9
Inception(v) [40]	30.33M	33.73G	79.49	38.91M	37.52G	24.6	43.1	25.6
RFB [10]	28.87M	33.72G	79.46	38.27M	37.15G	24.6	43.0	25.0
Ours (E-RFB)	28.74M	34.00G	**79.95**	35.57M	36.97G	**25.6**	**43.2**	**26.5**

TABLE V
COMPARISON OF DIFFERENT ATTENTION METHODS ON VOC 2007 TEST SET IN TERMS OF NETWORK PARAMETERS AND DETECTION ACCURACY. ΔPARAM. INDICATES THE ADDITIONAL PARAMETER RATIO INTRODUCED BY ATTENTION MODULE.

Baseline	Method	mAP	ΔParam.(%)
VGG-16 + E-RFB	ECA [14]	80.19	**0.000084**
	SE [12]	80.06	0.44
	CBAM [13]	80.04	0.45
	Ours (GSAM)	**80.23**	0.09

We compare our E-RFB with Residual block, RFB, Inception module and ASPP module. For a fair comparison, the channels of the internal branches require proportional adjustment to ensure the similar parameter and calculation amount of all the models. Four different attention modules are evaluated on the VGG-based E-RFB Net.

- E-RFB performs best when different blocks have similar magnitude.
- GSAM is an extremely lightweight attention module and achieves the highest accuracy.

CONCLUSION

- A new multibranch feature extraction module with group spatial attention mechanism is proposed.
- The proposed module is lightweight and can be easily integrated into various existing network architectures for feature enhancement.
- Experiments on the MS COCO and PASCAL VOC datasets show their advancement.

Thanks