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
{zhangjiacheng, zhaozc, sufei}@bupt.edu.cn

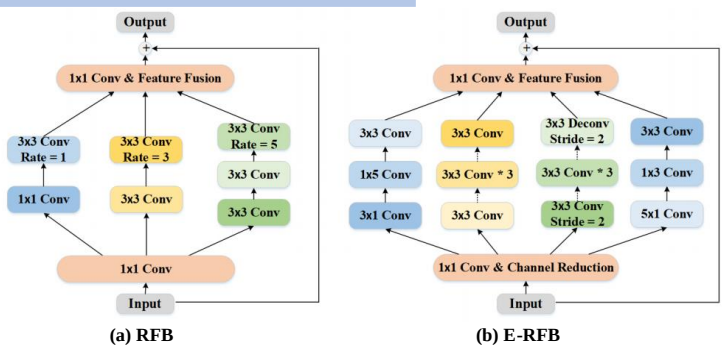
INTRODUCTION

We propose a novel multibranch feature extraction block - E-RFB, where sufficient RF is obtained by down-sampling and increasing depth.

In order to mitigate the spatial inconsistency in feature fusion, a novel spatial attention mechanism (GSAM) is proposed to model the internal relationship of a feature map.

The two aforementioned feature enhancement mechanisms are integrated in one module. The experimental results on the MS COCO and PASCAL VOC datasets demonstrate the effectiveness of the proposed module.

Efficient-Receptive Field Block

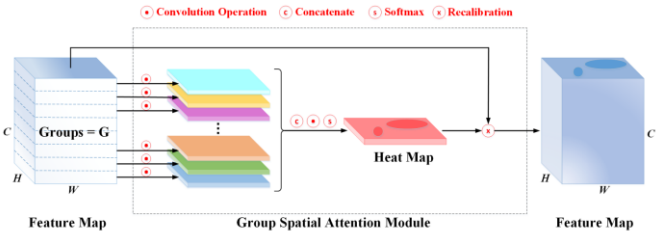


(a) RFB (b) E-RFB

E-RFB conjuncts output features of various levels to increase the richness of features.

- **RFs similar to RFB.** In the E-RFB, a large RF is obtained by down-sampling and depth growing rather than by dilated convolution.
- **Deep, narrow, and large cardinality.** A narrow structure creates less parameters, whereas a deep structure learns discriminative features.
- **Strip convolution.** To detect objects with extreme aspect ratios precisely, an n × m convolution implemented by adding an integrated n × 1 plus 1 × m conv layer is utilized.

Group Spatial Attention Module



(a) Structure of GSAM

(b) Visualization of heat maps

E-RFB adaptively learns the import degrees for different-level features on each channel by the following 1 × 1 convolution. Besides, we propose a novel spatial attention mechanism to eliminate the spatial inconsistency across different branches.

- To utilize the relationship within all response values comprehensively, we propose to gradually "narrow" the attention feature map. Meanwhile, for the sake of reducing computational burden, we divide the input feature map into multiple groups according to channel.
- There are three consecutive 3 × 3 convolution layers. Large RFs are crucial for spatial attention modeling.
- The ratio of additional parameter when the GSAM is added to ordinary k × k convolution layer is $r_s \approx 1/c_2$. Similarly, the additional FLOP ratio can be calculated as $r_s \approx 1/c_2 + 1/2 \cdot k^2 \cdot c_1$. The lightweight GSAM can be embedded easily into any network.

EXPERIMENTS

Our detectors surpassing the baseline SSD by a large margin regardless of the backbone. Compared with the RFB, the E-RFB contribute substantially to performance improvements in terms of AP_{small} and AP_{medium}. Benefiting from its large RF, GSAM captures context information in a large range, making for large object detection.

Compared with RFB Net, E-RFB Net has higher accuracy with fewer parameters, which confirms the efficiency of our module.

The E-RFB performs best when different blocks have similar magnitude, i.e. they have similar parameter size and FLOPs. Our module structure design contributes to detection accuracy.

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 [1]	12.38M	2.21G	20.0	35.5	20.2	1.9	19.6	36.2	4.0	30.7	52.1
	RFB [2]	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 [1]	50.98M	42.65G	29.2	48.2	30.8	11.2	31.5	44.5	16.3	44.7	59.2
	RFB [2]	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 [1]	45.81M	42.26G	32.3	51.7	34.6	14.7	35.6	45.6	21.2	50.0	62.1
	RFB [2]	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

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 [3]	**27.63M**	**32.73G**	78.74	**35.20M**	**36.52G**	24.4	42.6	24.8
ASPP(v) [4]	29.80M	33.78G	79.53	38.37M	37.57G	24.1	42.5	24.3
Inception [5]	29.18M	33.35G	78.86	37.76M	37.14G	24.3	42.0	24.9
Inception(v) [5]	30.33M	33.73G	79.49	38.91M	37.52G	24.6	43.1	25.6
RFB [2]	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**

REFERENCES/ACKNOWLEDGMENT

[1] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in European conference on computer vision. Springer, 2016, pp. 21–37.
[2] S. Liu, D. Huang, and a. Wang, "Receptive Field Block Net for Accurate and Fast Object Detection," in The European Conference on Computer Vision (ECCV), September 2018.
[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[4] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," arXiv preprint arXiv:1706.05587, 2017.
[5] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in Thirty-first AAAI conference on artificial intelligence, 2017.
This work is supported by Science and Technology Foundationof Beijing Municipal Science & Technology Commission(Z201100007520001), and Chinese National Natural ScienceFoundation (62076033 and U1931202)