CBAM: Convolutional Block Attention Module

Sanghyun Woo , Jongchan Park , Joon-Young Lee , and In So Kweon

ECCV 2018

Presenter: Minho Park

Contribution

1. We propose a simple yet effective attention module (CBAM) that can be widely applied to boost representation power of CNNs.
2. We validate the effectiveness of our attention module through extensive ablation studies.
3. We verify that performance of various networks is greatly improved on the multiple benchmarks (ImageNet-1K, MS COCO, and VOC 2007) by plugging our light-weight module.

The overview of CBAM

- The module has two sequential sub-modules: channel and spatial.

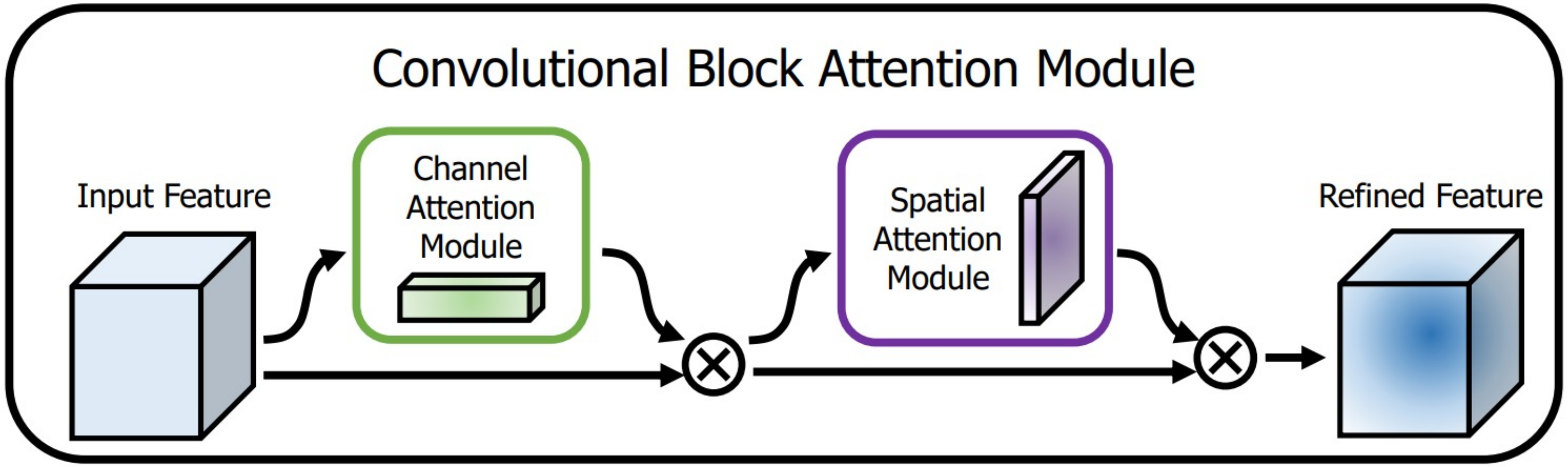

Fig. 1: The overview of CBAM.

The overview of CBAM

- Each of the branches can learn 'what' and 'where' to attend in the channel and spatial axes, respectively.
- As a result, our module efficiently helps the information flow within the network by learning which information to emphasize or suppress.

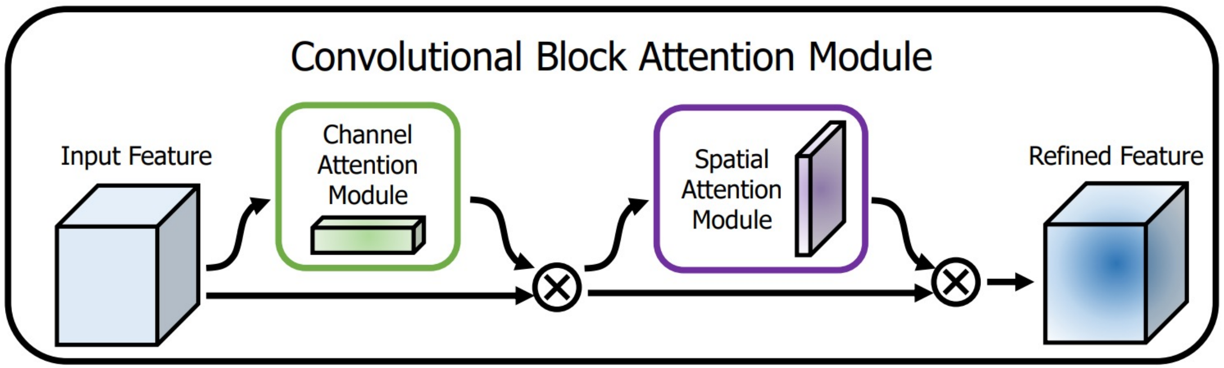

Fig. 1: The overview of CBAM.

Comparison

- Residual Attention Network for Image Classification, Fei Wang et al., CVPR 2017
- 3D Attention map (Computationally expensive)

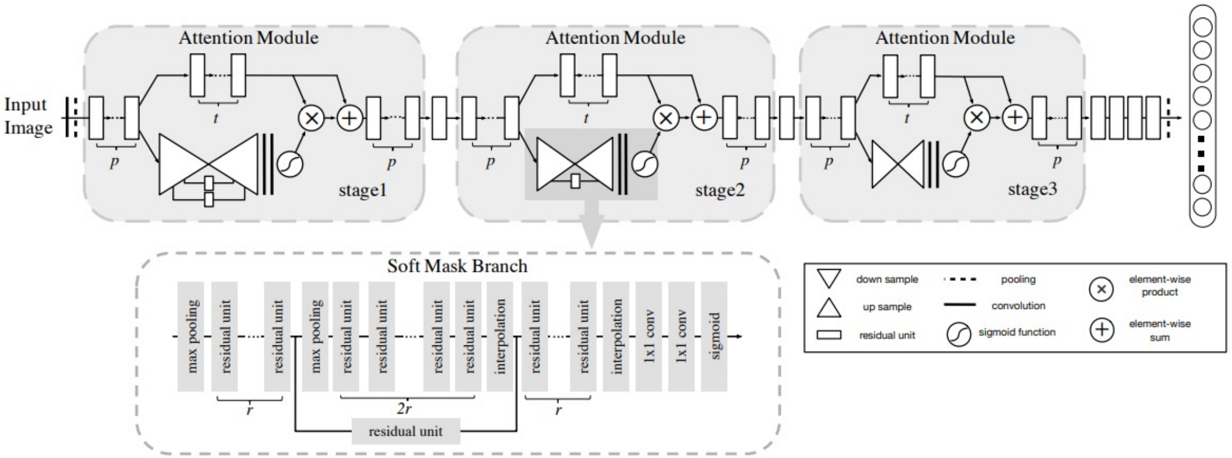

Fig. 2: Example architecture of the proposed network for ImageNet.

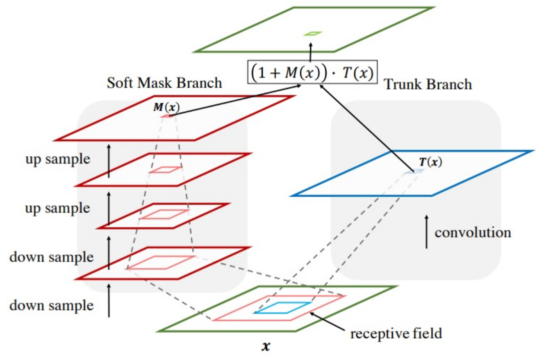

Fig. 3: The receptive field comparison between mask branch and trunk branch.

Comparison

- Squeeze-and-Excitation Networks, Jie Hu et al., CVPR 2018
- Only channel-wise attention

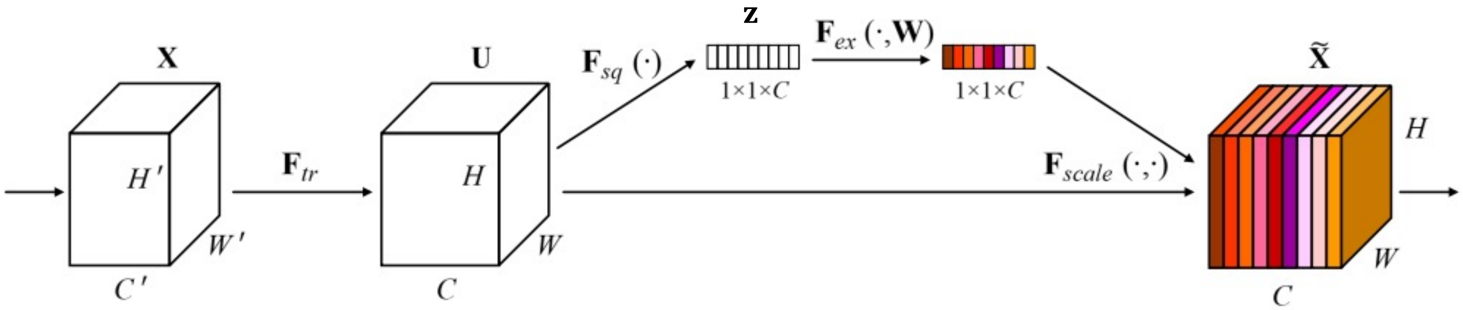

Fig. 2: A Squeeze-and-Excitation block.

- $\mathbf{F_{sq}}(\mathbf{U}) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} u_c(i,j)$, $\mathbf{F_{ex}}(\mathbf{z}, \mathbf{W}) = \sigma\big(\mathbf{W}_2 \mathrm{ReLU}(\mathbf{W}_1 \boldsymbol{z})\big)$

Method

- $\mathbf{M_c}(\mathbf{F}) = \sigma\left(\mathrm{MLP}\left(\mathrm{AvgPool_c}(\mathbf{F})\right) + \mathrm{MLP}\left(\mathrm{MaxPool_c}(\mathbf{F})\right)\right)$
- $\mathbf{M_s}(\mathbf{F}') = \sigma\left(f^{7\times 7}([\mathrm{AvgPool_s}(\mathbf{F}'); \mathrm{MaxPool_s}(\mathbf{F}')])\right)$

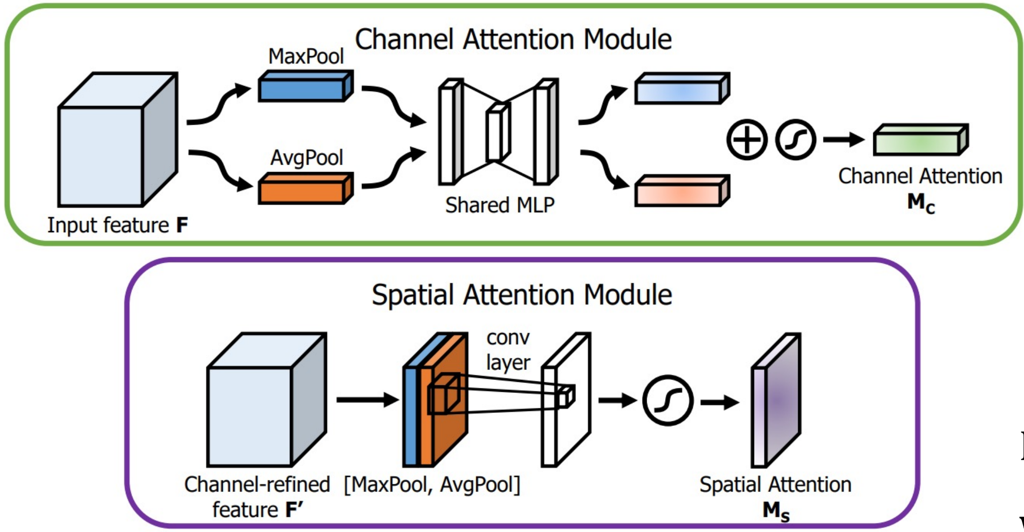

Fig. 2: Diagram of each attention sub-module.

$\mathrm{MLP}(\cdot)$: $\mathbf{W_1}\left(\mathrm{ReLU}\left(\mathbf{W_0}(\cdot)\right)\right)$

$\mathbf{W_0} \in \mathbb{R}^{\frac{C}{r}\times C}, \mathbf{W_1} \in \mathbb{R}^{C\times\frac{C}{r}}$

Method

- ResBlock + CBAM
- $\mathbf{F}' = \mathbf{M_c}(\mathbf{F}) \otimes \mathbf{F},\ \mathbf{F}'' = \mathbf{M_S}(\mathbf{F}') \otimes \mathbf{F}'$

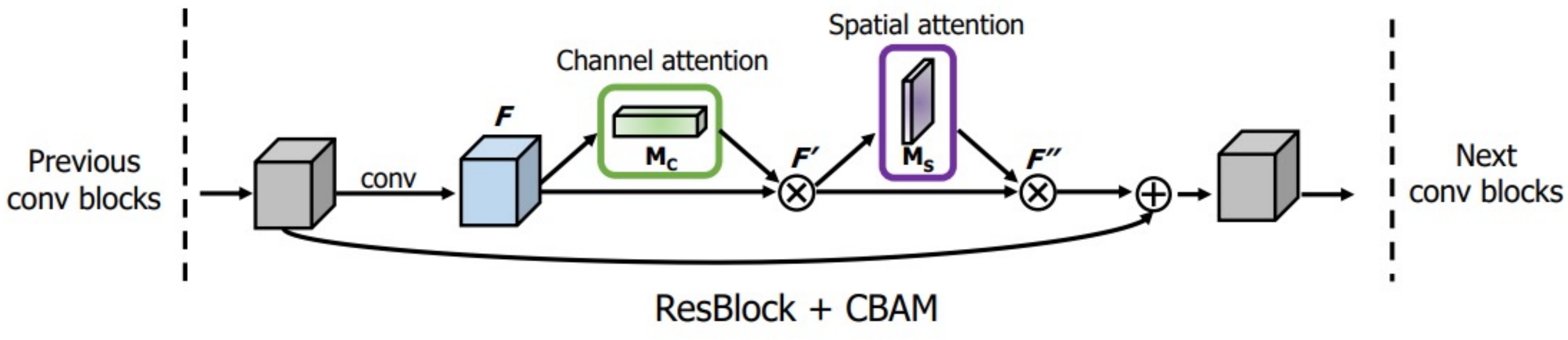

Fig. 3: CBAM integrated with a ResBlock in ResNet.

Experiments

- ImageNet-1K for image classification
- MS COCO and VOC 2007 for object detection

Ablation studies

- Channel attention: Verify that using both average-pooled and max-pooled features enables finer attention inference.
- AvgPool encodes global statistics softly.
- MaxPool encodes the degree of the most salient part.

Description	Parameters	GFLOPs	Top-1 Error(%)	Top-5 Error(%)
ResNet50 (baseline)	25.56M	3.86	24.56	7.50
ResNet50 + AvgPool (SE [28])	25.92M	3.94	23.14	6.70
ResNet50 + MaxPool	25.92M	3.94	23.20	6.83
ResNet50 + AvgPool & MaxPool	25.92M	4.02	**22.80**	**6.52**

Table 1: Comparison of different channel attention methods.

Ablation studies

- Spatial attention: Channel pooling (Max, Avg) is better than standard 1X1 convolution, and larger kernel size generates better accuracy.

Description	Param.	GFLOPs	Top-1 Error(%)	Top-5 Error(%)
ResNet50 + channel (SE [28])	28.09M	3.860	23.14	6.70
ResNet50 + channel	28.09M	3.860	22.80	6.52
ResNet50 + channel + spatial (1x1 conv, k=3)	28.10M	3.862	22.96	6.64
ResNet50 + channel + spatial (1x1 conv, k=7)	28.10M	3.869	22.90	6.47
ResNet50 + channel + spatial (avg&max, k=3)	28.09M	3.863	22.68	6.41
ResNet50 + channel + spatial (avg&max, k=7)	28.09M	3.864	**22.66**	**6.31**

Table 2: Comparison of different spatial attention methods.

Ablation studies

- Arrangement of the channel and spatial attention.
  - Parallel builds 3D attention map. (The outputs of the two attention modules are added and normalized with the sigmoid function.)

Description	Top-1 Error(%)	Top-5 Error(%)
ResNet50 + channel (SE [28])	23.14	6.70
ResNet50 + channel + spatial	**22.66**	**6.31**
ResNet50 + spatial + channel	22.78	6.42
ResNet50 + channel & spatial in parallel	22.95	6.59

Table 3: Combining methods of channel and spatial attention.

Image Classification on ImageNet-1K

- SOTA

Architecture	Param.	GFLOPs	Top-1 Error (%)	Top-5 Error (%)
ResNet18 [5]	11.69M	1.814	29.60	10.55
ResNet18 [5] + SE [28]	11.78M	1.814	29.41	10.22
ResNet18 [5] + CBAM	11.78M	1.815	**29.27**	**10.09**
ResNet34 [5]	21.80M	3.664	26.69	8.60
ResNet34 [5] + SE [28]	21.96M	3.664	26.13	8.35
ResNet34 [5] + CBAM	21.96M	3.665	**25.99**	**8.24**
ResNet50 [5]	25.56M	3.858	24.56	7.50
ResNet50 [5] + SE [28]	28.09M	3.860	23.14	6.70
ResNet50 [5] + CBAM	28.09M	3.864	**22.66**	**6.31**
ResNet101 [5]	44.55M	7.570	23.38	6.88
ResNet101 [5] + SE [28]	49.33M	7.575	22.35	6.19
ResNet101 [5] + CBAM	49.33M	7.581	**21.51**	**5.69**
WideResNet18 [6] (widen=1.5)	25.88M	3.866	26.85	8.88
WideResNet18 [6] (widen=1.5) + SE [28]	26.07M	3.867	26.21	8.47
WideResNet18 [6] (widen=1.5) + CBAM	26.08M	3.868	**26.10**	**8.43**
WideResNet18 [6] (widen=2.0)	45.62M	6.696	25.63	8.20
WideResNet18 [6] (widen=2.0) + SE [28]	45.97M	6.696	24.93	7.65
WideResNet18 [6] (widen=2.0) + CBAM	45.97M	6.697	**24.84**	**7.63**
ResNeXt50 [7] (32x4d)	25.03M	3.768	22.85	6.48
ResNeXt50 [7] (32x4d) + SE [28]	27.56M	3.771	**21.91**	6.04
ResNeXt50 [7] (32x4d) + CBAM	27.56M	3.774	21.92	**5.91**
ResNeXt101 [7] (32x4d)	44.18M	7.508	21.54	5.75
ResNeXt101 [7] (32x4d) + SE [28]	48.96M	7.512	21.17	5.66
ResNeXt101 [7] (32x4d) + CBAM	48.96M	7.519	**21.07**	**5.59**

Table 4: Classification results on ImageNet-1K.

Network Visualization with Grad-CAM

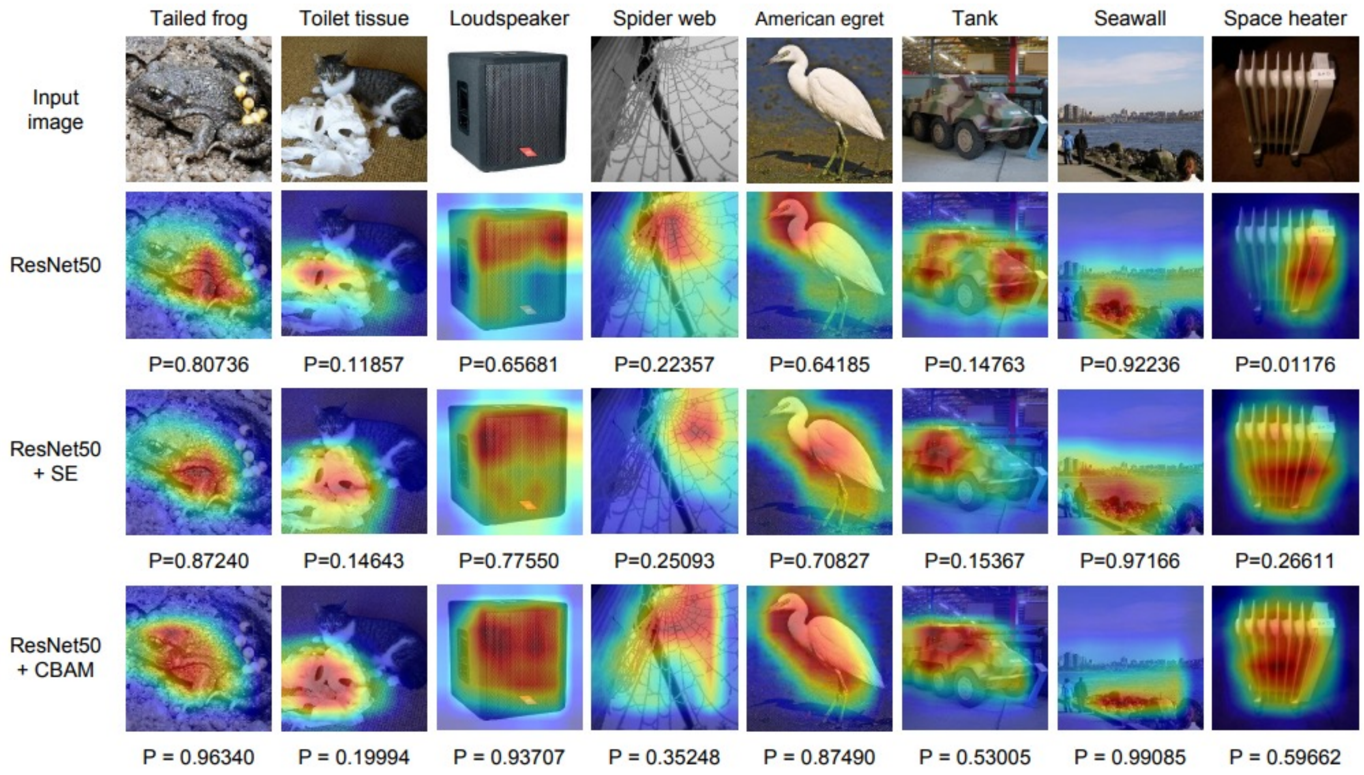

Fig. 5: Grad-CAM visualization results.

Network Visualization with Grad-CAM

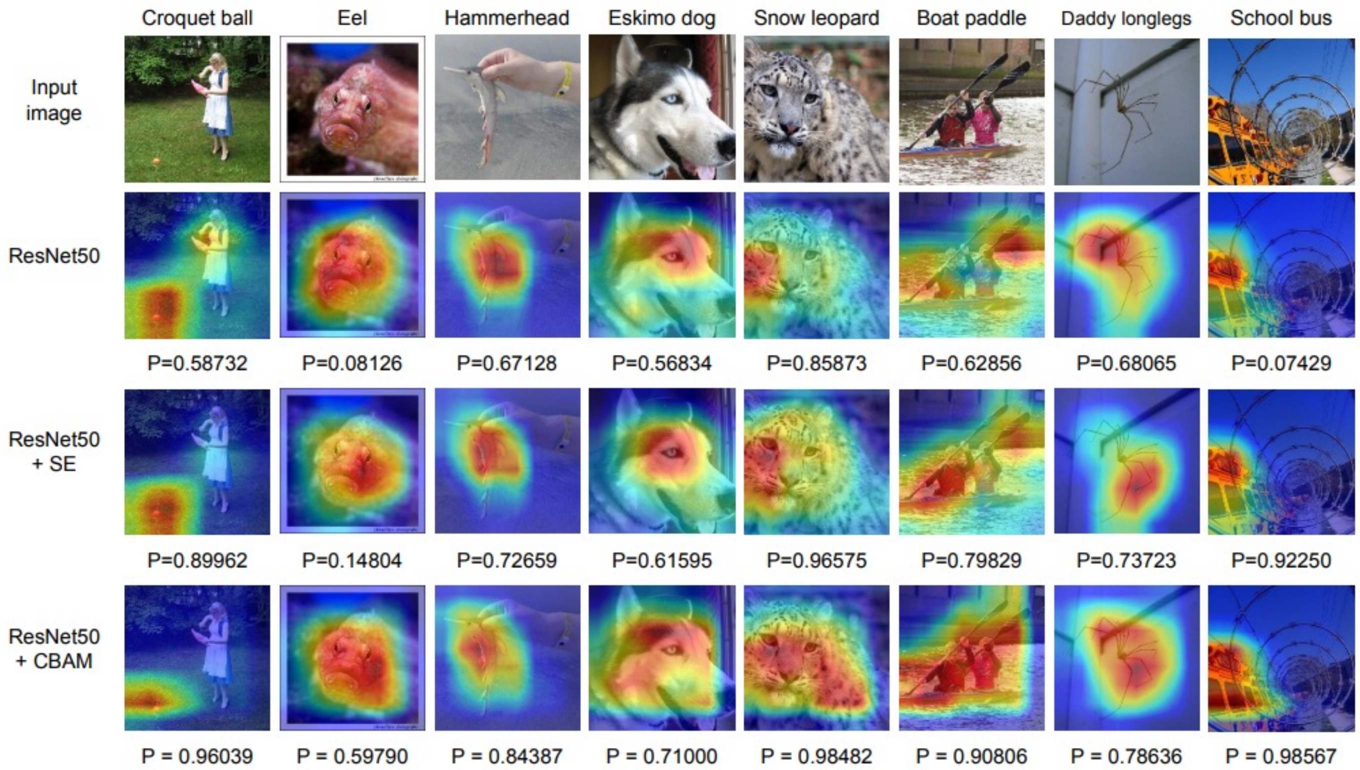

Fig. 5: Grad-CAM visualization results.

MS COCO Object Detection

- SOTA

Backbone	Detector	mAP@.5	mAP@.75	mAP@[.5, .95]
ResNet50 [5]	Faster-RCNN [41]	46.2	28.1	27.0
ResNet50 [5] + CBAM	Faster-RCNN [41]	**48.2**	**29.2**	**28.1**
ResNet101 [5]	Faster-RCNN [41]	48.4	30.7	29.1
ResNet101 [5] + CBAM	Faster-RCNN [41]	**50.5**	**32.6**	**30.8**

Table 6: Object detection mAP(%) on the MS COCO validation set.

VOC 2007 Object Detection

- SOTA

Backbone	Detector	mAP@.5	Parameters (M)
VGG16 [9]	SSD [39]	77.8	26.5
VGG16 [9]	StairNet [30]	78.9	32.0
VGG16 [9]	StairNet [30] + SE [28]	79.1	32.1
VGG16 [9]	StairNet [30] + CBAM	**79.3**	32.1
MobileNet [34]	SSD [39]	68.1	5.81
MobileNet [34]	StairNet [30]	70.1	5.98
MobileNet [34]	StairNet [30] + SE [28]	70.0	5.99
MobileNet [34]	StairNet [30] + CBAM	**70.5**	6.00

Table 7: Object detection mAP(%) on the VOC 2007 test set.

Thank you