EPSANet: An Efficient Pyramid Squeeze Attention Block on Convolutional Neural Network

Hu Zhang
VRLab, Beihang University, China
huzhang198@gmail.com

Keke ZU
Shenzhen University, Shenzhen, China
kekezu@szu.edu.cn

Jian Lu
Shenzhen University, Shenzhen, China
jianlu@szu.edu.cn

Yuru Zou
Shenzhen University, Shenzhen, China
yuruzou@szu.edu.cn

Deyu Meng
Xi'an Jiaotong University, Xi'an, China
dymeng@mail.xjtu.edu.cn

Citations 115

Table of Contents

1. Intorduction

- Attention mechanisms are widely used in many computer vision areas such as image classification, object detection, instance segmentation, semantic segmentation, scene parsing and action localization

- Two types of attention methods : Channel attention, Spatial attention .
- The most commonly used method of channel attention is the Squeeze-andExcitation (SE) module [13].
  - Advantage : low cost.
  - Disadvantage : Ignores the importance of spatial information
  - => Bottleneck Attention Module(BAM) [14]
  - => Convolutional Block Attention Module(CBAM) [5]

[13] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[14] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon. Bam: Bottleneck attention module. In British Machine Vision Conference(BMVC), 2018.
[5] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In European Conference on Computer Vision (ECCV), 2018.

1. Intorduction

- However, there still exists two important and challenging problems.
  1. How to efficiently capture and exploit the spatial information of the feature map with different scales to enrich the feature space.
  2. Channel or spatial attention can only effectively capture the local information but fail in establishing a long-range channel dependency.
     - => PyConv (Pyramid Convolution), Res2Net, HS-ResNet (Hierarchical Saliency Residual Network)

     * Those models have higher model complexity and thus the network suffers from heavy computational.

[15] Ionut Cosmin Duta, Li Liu, Fan Zhu, and Ling Shao. Pyramidal convolution:rethinking convolutional neural networks for visual recognition. arXiv preprint arXiv:2006.11538, 2020.
[16] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(2):652–662, 2021.
[17] Pengcheng Yuan, Shufei Lin, Cheng Cui, Yuning Du, Ruoyu Guo, Dongliang He, Errui Ding, and Shumin Han. Hs-resnet: Hierarchical-split block on convolutional neural network. arXiv preprint arXiv:2010.07621, 2020.

1. Intorduction

- Proposed => Pyramid Squeeze Attention (PSA)
  - PSA module has the ability to process the input tensor at multiple scales. Specifically, by using the multi-scale pyramid convolution structure to integrate the information of the input feature map.
    - **multi-scale pyramid convolution structure** => Helps take into account different object sizes or different spatial characteristics of features
    - **Channel-level compression** => Spatial information at different scales can be effectively extracted from the feature map for each channel.
    - Finally, the attention weight for each channel of the multi-scale feature map is obtained, the importance of each channel is normalized through Softmax, and a high weight is assigned to important information.

2. Related Work

1. Squeeze-and-Excitation (SE) attention [13] : capture channel correlations by selectively modulating the scale of channel.
2. Convolutional block attention module (CBAM) [5] : Enrich the attention map by adding max pooled features for the channel attention.
3. Selective kernel networks (SKNet) [21] : a dynamic selection attention mechanism that allows each neuron to adaptively adjust its receptive field size based on multiple scales of input feature map.
4. Efficient channel attention for deep convolutional neural networks (ECANet) [11] : one-dimensional convolution layer to reduce the redundancy of fully connected layers.
5. Dual attention network (DANet) [18] : adaptively integrates local features with their global dependencies by summing these two attention modules from different branches.

[13] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[5] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In European Conference on Computer Vision (ECCV), 2018.
[21] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 510–519, 2019.
[11] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.
[18] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene seg

3. Method

Channel attention

$X \in \mathbb{R}^{C \times H \times W}$ * H, W, C => height, width, number of input channels

$$g_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i,j)$$

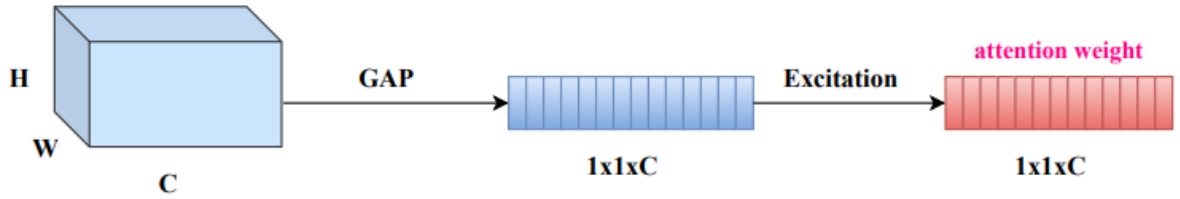


Figure 2: SEWeight module.

SE block consists of two parts :

squeeze and excitation

which is respectively designed for encoding the global information and adaptively recalibrating the channel-wise relationship.

3. Method

- **Original Channel attention**

$$g_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i,j)$$

$$w_c = \sigma(W_1 \delta(W_0(g_c)))$$

δ : Recitified Linear Unit (ReLU)

GAP : Global Average Pooling

σ : Excitation Function

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

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

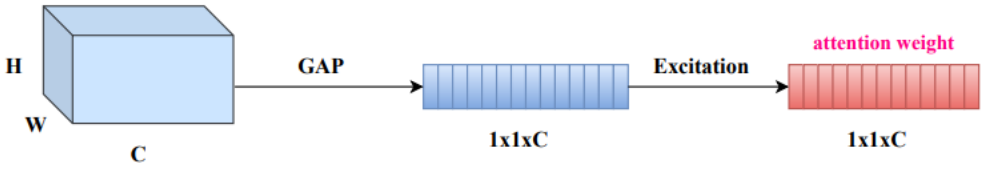
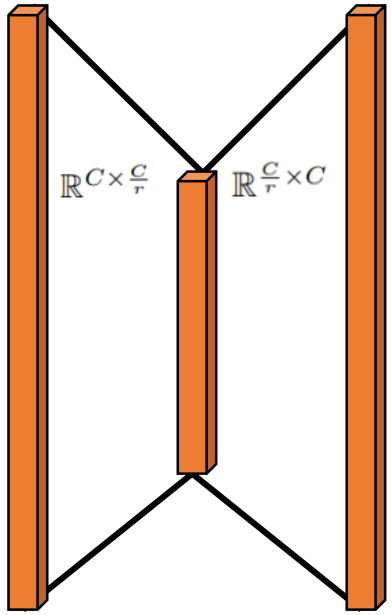


Figure 2: SEWeight module.



3. Method

- **Pyramid Squeeze Attention (PSA)**

1. First, the multi-scale feature map on channel-wise is obtained by implementing the proposed Squeeze and Concat (SPC) module.
2. Second, the channel-wise attention vector are obtained by using the SEWeight module to extract the attention of the feature map with different scales.
3. Third, re-calibrated weight of multi-scale channel is obtained by using the Softmax to re-calibrate the channel-wise attention vector.
4. Fourth, the operation of an element-wise product is applied to the re-calibrated weight and the corresponding feature map.

Finally, a refined feature map which is richer in multi-scale feature information can be obtained as the output.

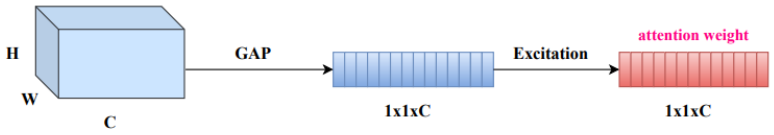


Figure 2: SEWeight module.

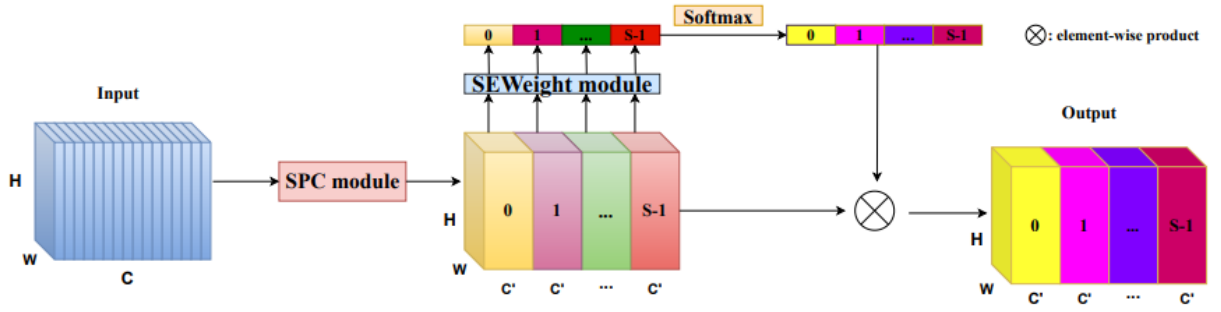


Figure 3: The structure of the proposed Pyramid Squeeze Attention (PSA) module.

3. Method

- **Pyramid Squeeze Attention (PSA)**

$$F_i = \mathrm{Conv}(k_i \times k_i, G_i)(X) \quad i = 0, 1, 2 \cdots S-1$$

K is the kernel size i-th kernel size $k_i = 2 \times (i+1) + 1$

G is the group size $G = 2^{\frac{K-1}{2}}$

$F_i \in R^{C' \times H \times W}$ $C' = \frac{C}{S}$

$$Z_i = \mathrm{SEWeight}(F_i), \quad i = 0, 1, 2 \cdots S-1 \quad Z_i \in R^{C' \times 1 \times 1}$$

$$Z = Z_0 \oplus Z_1 \oplus \cdots \oplus Z_{S-1}$$

$\oplus$: concat

Z_i is the attention value from the F_i

Z is the multi-scale attention weight vector.

$$F = \mathrm{Cat}([F_0, F_1, \cdots, F_{S-1}]) \tag{5}$$

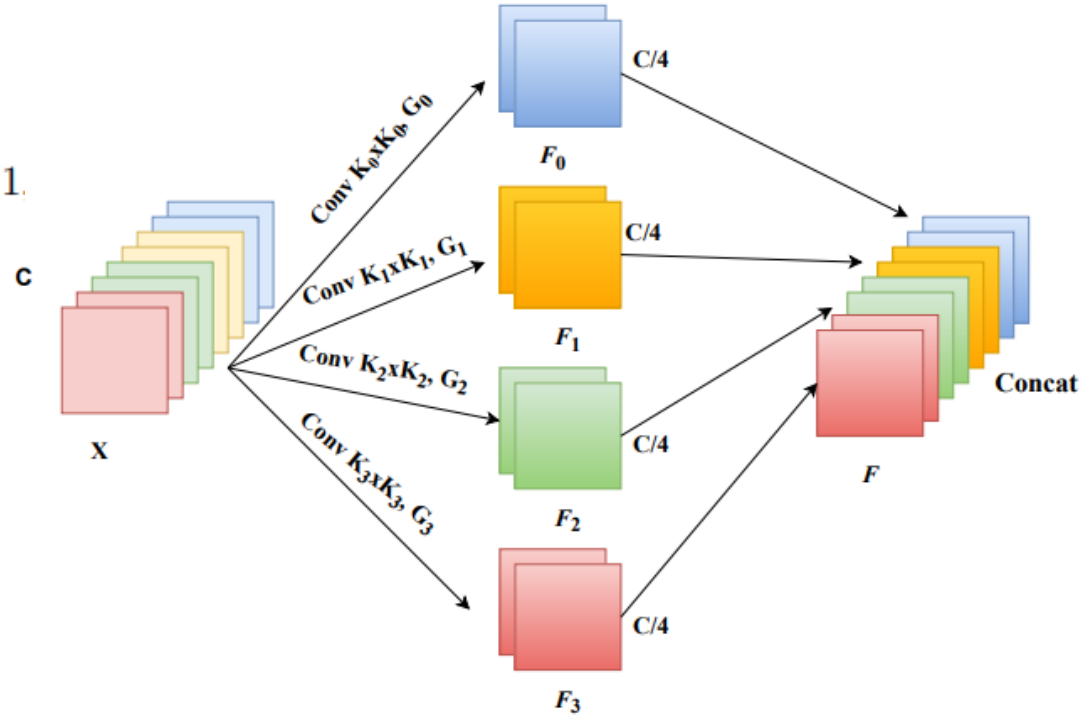


Figure 4: A detailed illustration of the proposed Squeeze and Concat (SPC) module with S=4, where 'Squeeze' means to equally squeeze in the channel dimension, K is the kernel size, G is the group size and 'Concat' means to concatenate features in the channel dimension.

Extract the spatial information of the input feature map in a multi-branch way

3. Method

- **Pyramid Squeeze Attention (PSA)**

$$att_i = \mathrm{Softmax}(Z_i) = \frac{exp(Z_i)}{\sum_{i=0}^{S-1} exp(Z_i)}$$

$$att = att_0 \oplus att_1 \oplus \cdots \oplus att_{S-1}$$

$$Y_i = F_i \odot att_i \quad i = 1, 2, 3, \cdots S-1$$

$\odot$ represents the channel-wise multiplication

$$Out = \mathrm{Cat}([Y_0, Y_1, \cdots, Y_{S-1}])$$

$$F = \mathrm{Cat}([F_0, F_1, \cdots, F_{S-1}]) \tag{5}$$

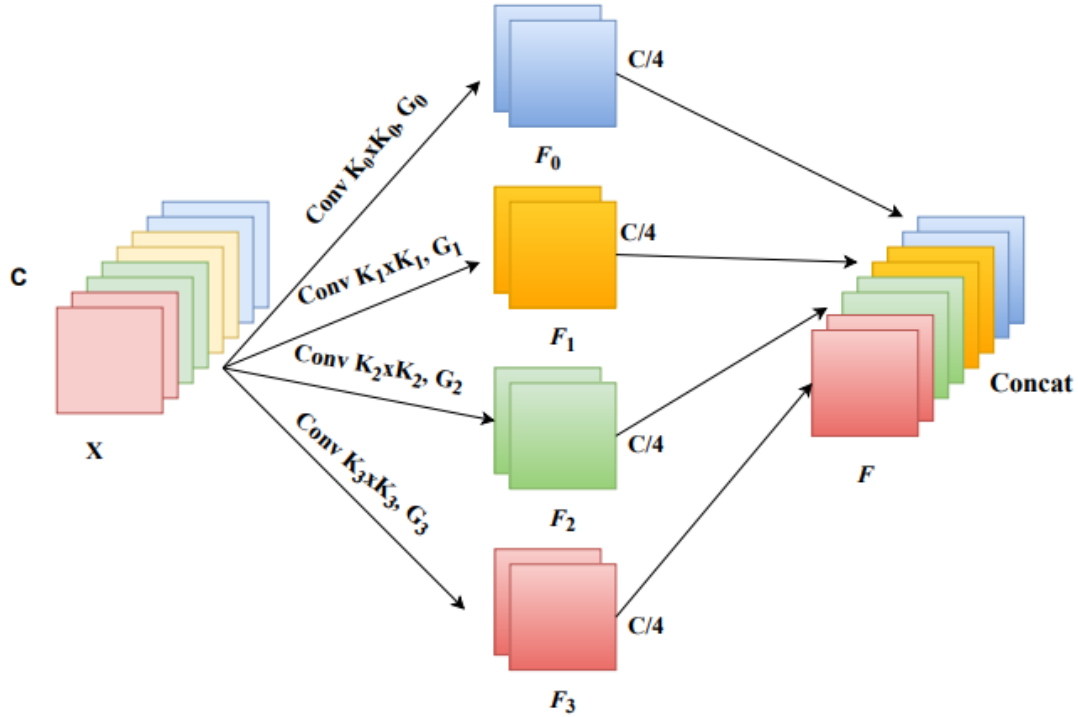


Figure 4: A detailed illustration of the proposed Squeeze and Concat (SPC) module with S=4, where 'Squeeze' means to equally squeeze in the channel dimension, K is the kernel size, G is the group size and 'Concat' means to concatenate features in the channel dimension.

3. Method

- **Pyramid Squeeze Attention (PSA)**

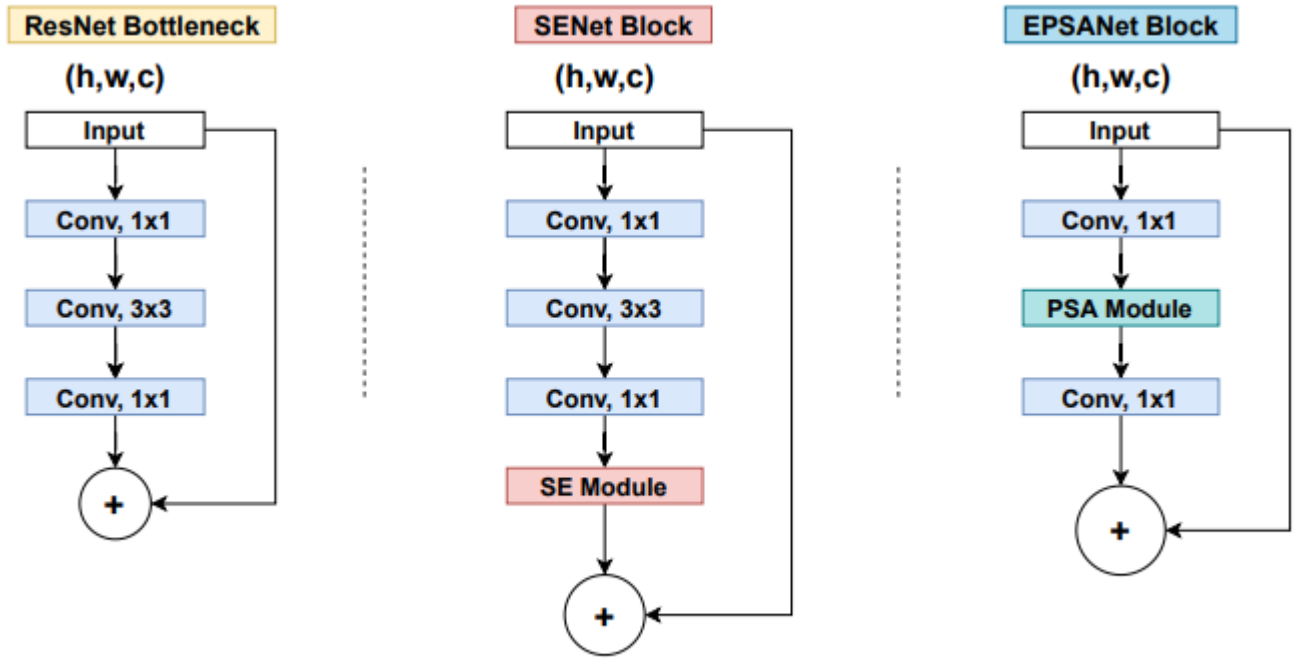


Figure 5: Illustration and comparison of ResNet, SENet, and our proposed EPSANet blocks.

3. Method

- **Pyramid Squeeze Attention (PSA)**

Table 1: Network design of the proposed EPSANet.

Output	ResNet-50	EPSANet-50(Small)	EPSANet-50(Large)
112×112	7×7, 64, stride 2		
56×56	3×3 max pool, stride 2		
56×56	$\begin{bmatrix} 1\times1, & 64 \\ 3\times3, & 64 \\ 1\times1, & 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, & 64 \\ PSA, & 64 \\ 1\times1, & 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, & 128 \\ PSA(G=32), & 128 \\ 1\times1, & 256 \end{bmatrix}\times3$
28×28	$\begin{bmatrix} 1\times1, & 128 \\ 3\times3, & 128 \\ 1\times1, & 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, & 128 \\ PSA, & 128 \\ 1\times1, & 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, & 256 \\ PSA(G=32), & 256 \\ 1\times1, & 512 \end{bmatrix}\times4$
14×14	$\begin{bmatrix} 1\times1, & 256 \\ 3\times3, & 256 \\ 1\times1, & 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, & 256 \\ PSA, & 256 \\ 1\times1, & 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, & 512 \\ PSA(G=32), & 512 \\ 1\times1, & 1024 \end{bmatrix}\times6$
7×7	$\begin{bmatrix} 1\times1, & 512 \\ 3\times3, & 512 \\ 1\times1, & 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, & 512 \\ PSA, & 512 \\ 1\times1, & 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, & 1024 \\ PSA(G=32), & 1024 \\ 1\times1, & 2048 \end{bmatrix}\times3$
1 × 1	7×7 global average pool,1000-d fc		

4. Experiments

Table 2: Comparison of various attention methods on ImageNet in terms of network parameters(in millions), floating point operations per second (FLOPs), Top-1 and Top-5 Validation Accuracy(%).

Network	Backbone	Parameters	FLOPs	Top-1 Acc. (%)	Top-5 Acc. (%)
ResNet[27]	ResNet-50	25.56	4.12G	75.20	92.52
SENet[13]		28.07	4.13G	76.71	93.38
CBAM[5]		28.07	4.14G	77.34	93.69
A^2-Net[8]		33.00	6.50G	77.00	93.50
ABN[28]		43.59	7.18G	76.90	-
GCNet[1]		28.11	4.13G	77.70	93.66
Triplet Attention[29]		25.56	4.17G	77.48	93.68
AANet[30]		25.80	4.15G	77.70	93.80
ECANet[11]		25.56	4.13G	77.48	93.68
FcaNet[9]		28.07	4.13G	78.52	94.14
EPSANet(Small)		**22.56**	**3.62G**	77.49	93.54
EPSANet(Large)		27.90	4.72G	**78.64**	**94.18**
ResNet[27]	ResNet-101	44.55	7.85G	76.83	93.48
SENet[13]		49.29	7.86G	77.62	93.93
BAM[14]		44.91	7.93G	77.56	93.71
CBAM[5]		49.33	7.88G	78.49	94.31
SRM[31]		44.68	7.95G	78.47	94.20
ECANet[11]		44.55	7.86G	78.65	94.34
AANet[30]		45.40	8.05G	78.70	94.40
Triplet Attention[29]		44.56	7.95G	78.03	93.85
EPSANet(Small)		**38.90**	**6.82G**	78.43	94.11
EPSANet(Large)		49.59	8.97G	**79.38**	**94.58**

- EPSANet(Small), the **kernel** and **group size** are respectively set as (3,5,7,9) and (1,4,8,16) in the SPC module.
- EPSANet(Large) has a higher **group size** and is set as (32,32,32,32) in the SPC module.

4. Experiments

Table 2: Comparison of various attention methods on ImageNet in terms of network parameters(in millions), floating point operations per second (FLOPs), Top-1 and Top-5 Validation Accuracy(%).

Network	Backbone	Parameters	FLOPs	Top-1 Acc. (%)	Top-5 Acc. (%)
ResNet[27]	ResNet-50	25.56	4.12G	75.20	92.52
SENet[13]		28.07	4.13G	76.71	93.38
CBAM[5]		28.07	4.14G	77.34	93.69
A^2-Net[8]		33.00	6.50G	77.00	93.50
ABN[28]		43.59	7.18G	76.90	-
GCNet[1]		28.11	4.13G	77.70	93.66
Triplet Attention[29]		25.56	4.17G	77.48	93.68
AANet[30]		25.80	4.15G	77.70	93.80
ECANet[11]		25.56	4.13G	77.48	93.68
FcaNet[9]		28.07	4.13G	78.52	94.14
EPSANet(Small)		**22.56**	**3.62G**	77.49	93.54
EPSANet(Large)		27.90	4.72G	**78.64**	**94.18**
ResNet[27]	ResNet-101	44.55	7.85G	76.83	93.48
SENet[13]		49.29	7.86G	77.62	93.93
BAM[14]		44.91	7.93G	77.56	93.71
CBAM[5]		49.33	7.88G	78.49	94.31
SRM[31]		44.68	7.95G	78.47	94.20
ECANet[11]		44.55	7.86G	78.65	94.34
AANet[30]		45.40	8.05G	78.70	94.40
Triplet Attention[29]		44.56	7.95G	78.03	93.85
EPSANet(Small)		**38.90**	**6.82G**	78.43	94.11
EPSANet(Large)		49.59	8.97G	**79.38**	**94.58**

- EPSANet(Small), the **kernel** and **group size** are respectively set as (3,5,7,9) and (1,4,8,16) in the SPC module.
- EPSANet(Large) has a higher **group size** and is set as (32,32,32,32) in the SPC module.

謝謝聆聽