SA-Net: Shuffle Attention for Deep Convolutional Neural Networks

Qing-Long Zhang, Yu-Bin Yang

State Key Laboratory for Novel Software Technology, Nanjing University, China

Codes and pretrained models are available at https://github.com/wofmanaf/SA-Net

Introduction

- **Attention Mechanism**
1. Correctly incorporating attention mechanisms into convolution blocks can significantly improve the performance of CNNs.
2. There are mainly two types of attention mechanisms most commonly used in computer vision: channel attention and spatial attention.
3. Integrated spatial attention and channel attention into one module can achieve significant improvement.
4. Existing works suffered from either converging difficulty or heavy computation burdens.

Shuffle Attention

- **Overall Architecture**

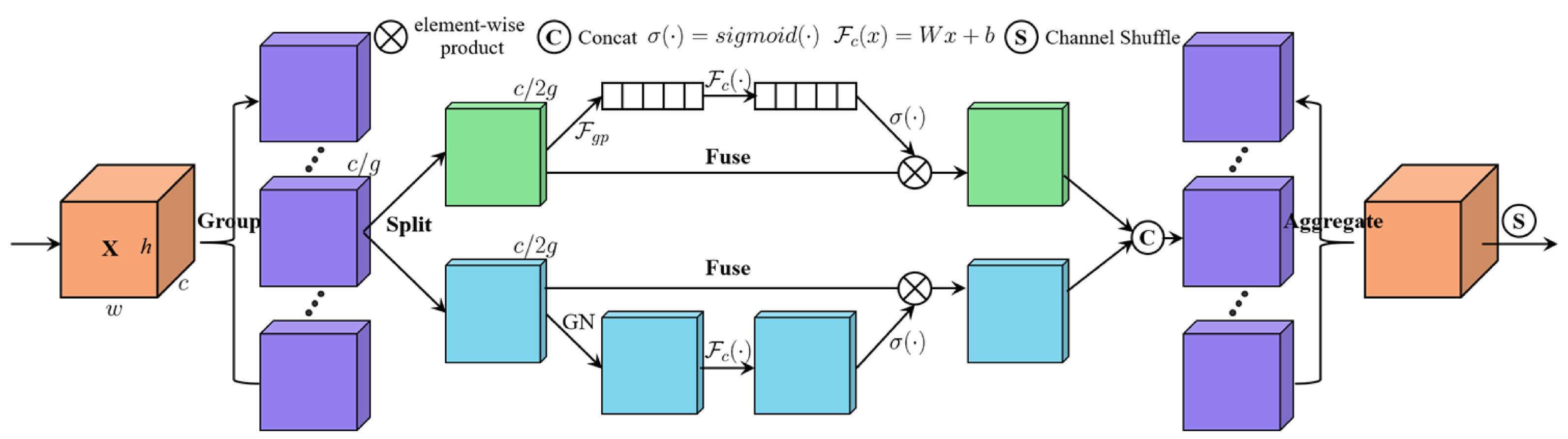


Figure 1: SA module divides the input feature map into groups, and uses Shuffle Unit to integrate the channel attention and spatial attention into one block for each group. After that, all sub-features are aggregated and a "channel shuffle" operator is utilized to enable information communication between different sub-features.

- **Feature Grouping**
  - For a given feature map $x \in \mathbb{R}^{c \times h \times w}$, SA first divides x into g groups along the channel dimension, i.e., $x = [x_1, \cdots, x_g]$, in which each sub-feature x_k gradually captures a specific semantic response in the training process.
  - Then, we generate the corresponding importance coefficient for each sub-feature through an attention module. Specifically, at the beginning of each attention unit, the input of x_k is split into two branches along the channels dimension.
  - one branch is adopted to produce a channel attention map by exploiting the inter-relationship of channels, while the other branch is used to generate a spatial attention map by utilizing the inter-spatial relationship of features, so that the model can focus on "what" and "where" is meaningful.

- **Channel Attention**
  - Using GAP to generate channel-wise statistics as $s \in \mathbb{R}^{c/2g \times 1 \times 1}$ by shrinking x_{k1} through spatial dimension, i.e.,

$$s = \mathcal{F}_{gp}(x_{k1}) = \frac{1}{h \times w} \sum_{i=1}^{h} \sum_{j=1}^{w} x_{k1}(i,j) \quad (1)$$

  - Creating a compact feature to enable guidance for adaptive selection, i.e.,

$$x'_{k1} = \sigma(\mathcal{F}_c(s)) \cdot x_{k1} = \sigma(W_1 s + b_1) \cdot x_{k1} \quad (2)$$

Where W_1 and b_1 are parameters used to scale and shift s.

Shuffle Attention

- **Spatial Attention**
  - Using Group Norm (GN) over x_{k2} to obtain spatial-wise statistics, i.e.,

$$\hat{x}_{k2} = \mathrm{GN}(x_{k2}) \quad (3)$$

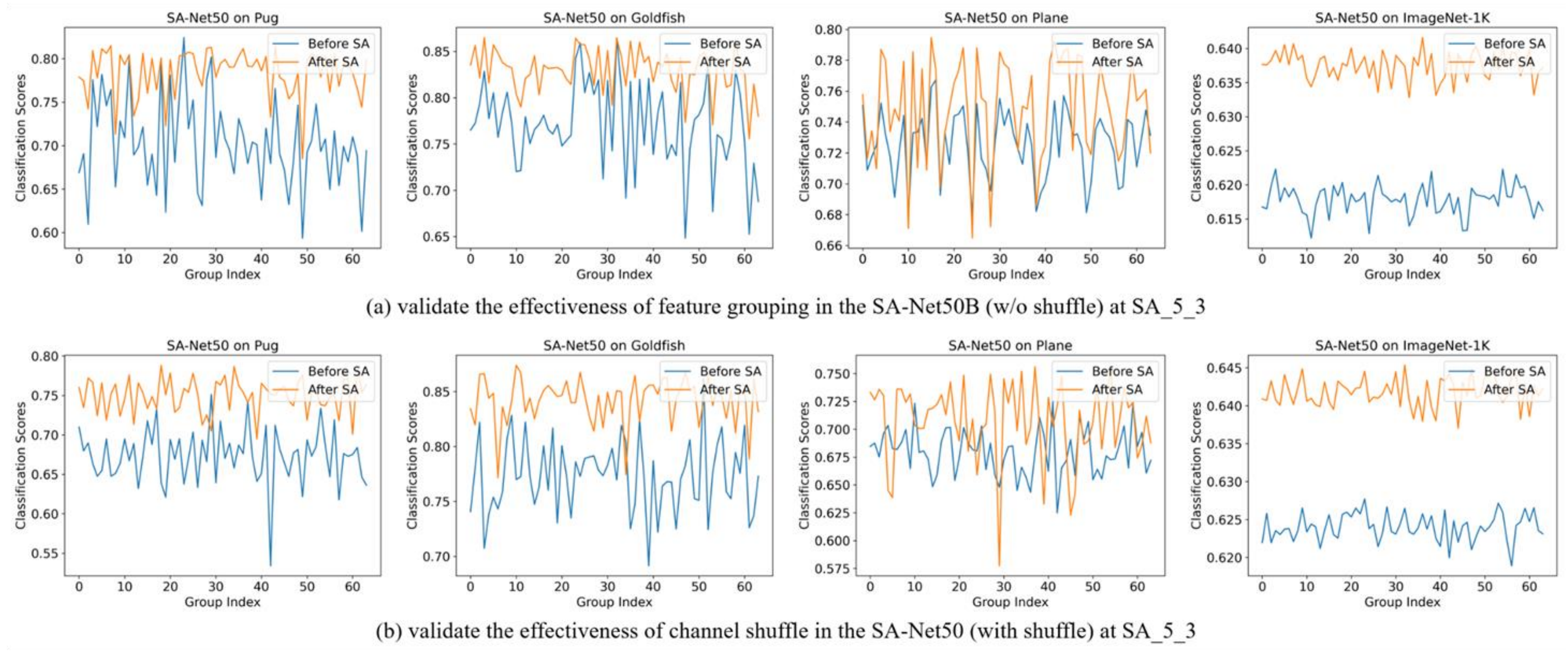
  - Adopting $\mathcal{F}_c$ to enhance the representation of $\hat{x}_{k2}$. The final output of spatial attention is obtained by

$$x'_{k2} = \sigma(W_2 \hat{x}_{k2} + b_2) \cdot x_{k2} \quad (4)$$

Where W_2 and b_2 are parameters used to scale and shift $\hat{x}_{k2}$.

Experiments

- **Classification on ImageNet-1k**



(a) validate the effectiveness of feature grouping in the SA-Net50B (w/o shuffle) at SA_5_3

(b) validate the effectiveness of channel shuffle in the SA-Net50 (with shuffle) at SA_5_3

Figure 2: Validation on the effectiveness of SA.

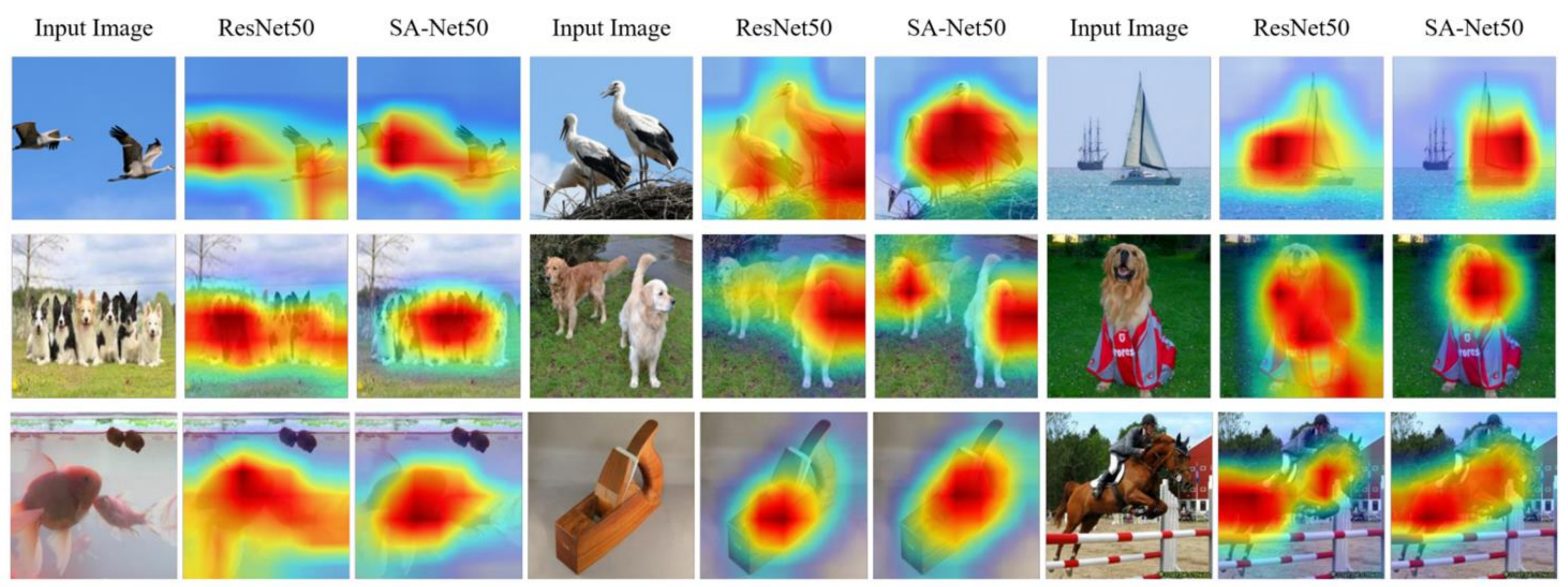


Figure 3: Sample visualization on ImageNet-1k validation split generated by GradCAM. All target layer selected is "layer4.2".

- **Ablation Studies**

Table 1: Ablation studies of SA-Net50 on ImageNet-1k dataset with four options (i.e., eliminating Group Norm, eliminating Channel Shuffle, eliminating $\mathcal{F}_c(\cdot)$ and utilizing Conv-1x1 to replace $\mathcal{F}_c(\cdot)$.

Methods	GFLOPs	Top-1 Acc (%)	Top-5 Acc (%)
origin	4.125	**77.724**	93.798
w/o_gn	4.125	77.372	93.804
w/o_shuffle	4.125	77.598	93.758
w/o_$\mathcal{F}_c(\cdot)$	4.125	77.608	**93.886**
1×1 Conv	4.140	77.684	93.840

Experiments

- **Classification on ImageNet-1k**

Table 2: Comparisons of different attention methods on ImageNet-1k in terms of network parameters (Param.), GFLOPs, and Top-1/Top-5 accuracy (in \%).

Attention Methods	Backbones	Param.	GFLOPs	Top-1 Acc (%)	Top-5 Acc (%)
ResNet [17]	ResNet-50	25.557M	4.122	76.384	92.908
SENet [23]		28.088M	4.130	77.462	93.696
CBAM [10]		28.090M	4.139	77.626	93.660
SGE-Net [12]		25.559M	4.127	77.584	93.664
ECA-Net [11]		**25.557M**	4.127	77.480	93.680
SA-Net (Ours)		**25.557M**	**4.125**	**77.724** (↑ 1.34)	**93.798** (↑ 0.89)
ResNet [17]	ResNet-101	44.549M	7.849	78.200	93.906
SENet [23]		49.327M	7.863	78.468	94.102
CBAM [10]		49.330M	7.879	78.354	94.064
SGE-Net [12]		44.553M	7.858	78.798	94.368
ECA-Net [11]		**44.549M**	7.858	78.650	94.340
SA-Net (Ours)		44.551M	**7.854**	**78.960** (↑ 0.76)	**94.492** (↑ 0.59)

- **Object Detection on MS COCO**

Table 3: Object detection results of different attention methods on COCO val2017.

Methods	Detectors	Param.	GFLOPs	AP50:95	AP50	AP75	AP_S	AP_M	AP_L
ResNet-50	Faster R-CNN	41.53M	207.07	36.4	58.4	39.1	21.5	40.0	46.6
+ SE		44.02M	207.18	37.7	60.1	40.9	22.9	41.9	48.2
+ SA (Ours)		**41.53M**	**207.35**	**38.7** (↑ 2.3)	**61.2**	**41.4**	**22.3**	**42.5**	**49.8**
ResNet-101		60.52M	283.14	38.5	60.3	41.6	22.3	43.0	49.8
+ SE		65.24M	283.33	39.6	62.0	43.1	23.7	44.0	51.4
+ SA (Ours)		**60.53M**	**283.60**	**41.0**(↑ 2.5)	**62.7**	**44.8**	**24.4**	**45.1**	**52.5**
ResNet-50	Mask R-CNN	44.18M	275.58	37.3	59.0	40.2	21.9	40.9	48.1
+ SE		46.67M	275.69	38.7	60.9	42.1	23.4	42.7	50.0
+ SA (Ours)		**44.18M**	**275.86**	**39.4**(↑ 2.1)	**61.5**	**42.6**	**23.4**	**42.8**	**51.1**
ResNet-101		63.17M	351.65	39.4	60.9	43.3	23.0	43.7	51.4
+ SE		67.89M	351.84	40.7	62.5	44.3	23.9	45.2	52.8
+ SA (Ours)		**63.17M**	**352.10**	**41.6**(↑ 2.2)	**63.0**	**45.5**	**24.9**	**45.5**	**54.2**
ResNet-50	RetinaNet	37.74M	239.32	35.6	55.5	38.3	20.0	39.6	46.8
+ SE		40.25M	239.43	36.0	56.7	38.3	20.5	39.7	**47.7**
+ SA (Ours)		**37.74M**	**239.60**	**37.5**(↑ 1.9)	**58.5**	**39.7**	**21.3**	**41.2**	45.9
ResNet-101		56.74M	315.39	37.7	57.5	40.4	21.1	42.2	49.5
+ SE		61.49M	315.58	38.8	59.3	41.7	22.1	43.2	51.5
+ SA (Ours)		**56.64M**	**315.85**	**40.3**(↑ 2.6)	**61.2**	**43.2**	**23.2**	**44.4**	**53.5**

- **Instance Segmentation on MS COCO.**

Table 4: Instance segmentation results of various state-of-the-arts attention modules using Mask R-CNN on COCO val2017.

Methods	AP50:95	AP50	AP75	AP_S	AP_M	AP_L
ResNet-50	34.2	55.9	36.2	18.2	37.5	46.3
+ SE	35.4	57.4	37.8	17.1	38.6	**51.8**
+ ECA	35.6	58.1	37.7	17.6	39.0	**51.8**
+ SGE	34.9	56.9	37.0	19.1	38.4	47.3
+ SA (Ours)	**36.1**(↑ 1.9)	**58.7**	**38.2**	**19.4**	**39.4**	49.0
ResNet-101	35.9	57.7	38.4	19.2	39.7	49.7
+ SE	36.8	59.3	39.2	17.2	40.3	53.6
+ ECA	37.4	59.9	39.8	18.1	41.1	**54.1**
+ SGE	36.9	59.3	39.4	20.0	40.8	50.1
+ SA (Ours)	**38.0**(↑ 2.1)	**60.0**	**40.3**	**20.8**	**41.2**	51.7

Conclusion

- ✓ We introduce SA module for deep CNNs, which groups channel dimensions into multiple sub-features, and then utilizes a Shuffle Unit to integrate the complementary channel and spatial attention module for each sub-feature.
- ✓ Extensive experimental results on ImageNet-1k and MS COCO demonstrate that the proposed SA has lower model complexity than the state-of-the-art attention approaches while achieving outstanding performance.