SCA Net: Sparse Channel Attention Module for Action Recognition

Hang Song*, YongHong Song, YuanLin Zhang

College of Artificial Intelligence - Xi'an Jiaotong University

Introduction

Our work aims at solving the following problem: How to learn channel attention with a more lightweight method? Thus, we develop our more lightweight channel attention module that can be easily incorporated into mainstream backbones for action recognition.

-We empirically demonstrate the importance of avoiding dense channel connection in learning channel attention, which motivates us to propose a more effective and efficient Sparse Channel Attention (SCA) module.

-We develop a novel Aggregate-Shuffle-Diverge (ASD) function to enhance local cross-group interaction, which brings no additional parameters but clearly improve the data representation capacity of SCA.

-We employ a multi-layer with small filters structure to accomplish 3D convolution factorization at the very beginning of our backbone network, so as to further reduce the size of the network.

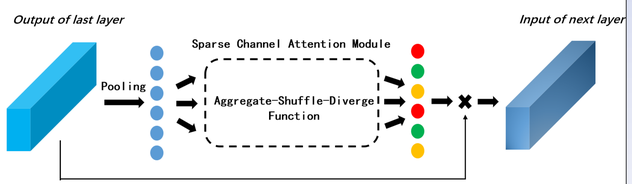

Figure 1: The overview of SCA module

Method

SCA module:
we denote our SCA module as:

$$Y = \delta(\sum_{i=1}^{G} F_{i(W_1,W_2)}(P(x_i)))$$

And, we denote our ASD function as:

$$F_{\omega_i} = W_2 \bigotimes \sigma(shuffle(\sum_{i=1}^{G} W_1 \bigotimes \omega_i))$$

Thus, our SCA module can be illustrated as the following picture:

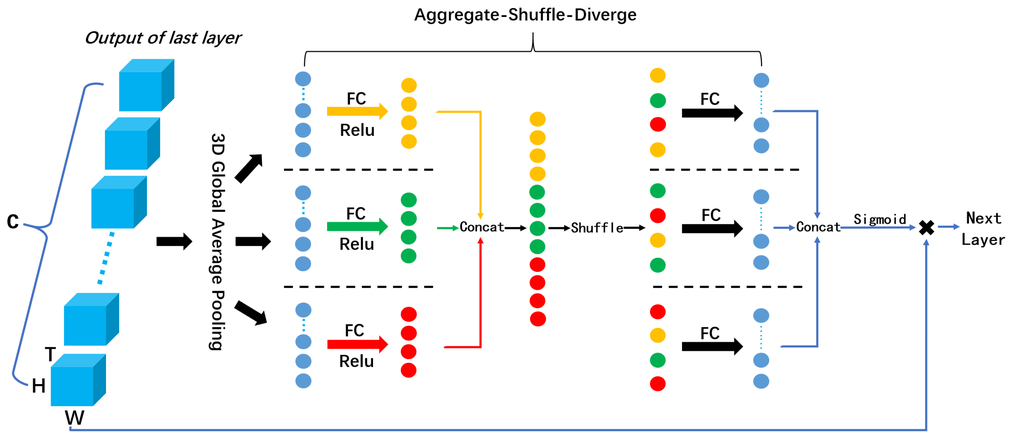

Figure 2: The specific architecture of Sparse Channel Attention (SCA) module

3D Convolution Factorization:

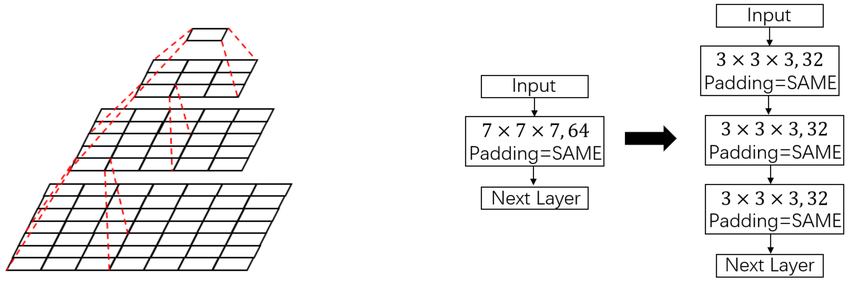

Figure 3(a) **Figure 3(b)**

Figure 3: Illustration of 3D Convolution Factorization.

(a) illustrates how we achieve a 7x7x7 receptive filed with three 3x3x3 convolutional filters. And (b) shows the specific structure we employ in our experiments.

Experiment results

Methods	Para.	Para.*	Top-1	Top-5
Factorized I3D	0	0	60.23%	84.56%
+SE-r_2	C^2	2.41M	63.94%	86.25%
+SE-r_4	$C^2/2$	1.20M	64.38%	86.55%
+SE-r_8	$C^2/4$	0.60M	63.26%	85.87%
+SE-r_{16}	$C^2/8$	0.30M	63.62%	85.25%
+SCA-G_2 (Ours)	$C^2/2$	1.20M	**67.22%**	**88.28%**
+SCA-G_4 (Ours)	$C^2/4$	0.60M	66.40%	88.18%
+SCA-G_8 (Ours)	$C^2/8$	0.30M	65.35%	87.59%
+SCA-G_{16} (Ours)	$\boldsymbol{C^2/16}$	**0.15M**	64.10%	86.76%

Table 1:
Comparison with SE block using factorized I3D as backbone network training from scratch on UCF-101.

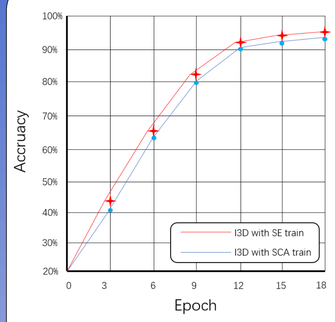

Figure 4:
Training accuracy as a function of training iterations for I3D with SE block and SCA module. The red line stands for training I3D with SE block, while the blue line stands for I3D with SCA module.

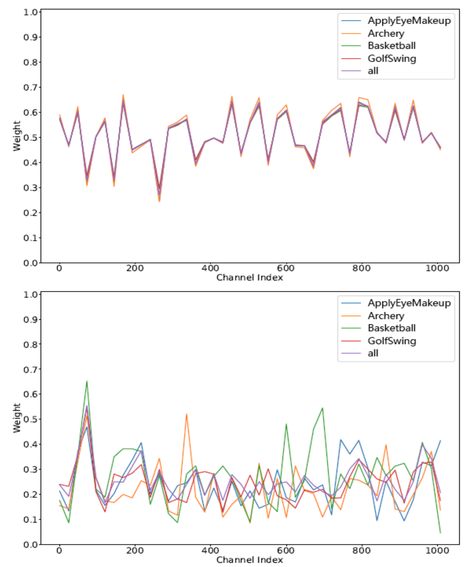

Figure 5:
Visualization the channel weights in 6'th block. The channel weights learned by SCA modules and SE blocks are illustrated in bottom and top.

Methods	Top-1	Top-5	Average
Slow Fusion [19]	41.32%	67.53%	63.94%
C3D [8]	44.08%	70.23%	64.23%
LTC [20]	48.41%	74.03%	63.26%
MicT-Net [21]	50.95%	76.21%	63.26%
SM-ConvLSTM [22]	54.28%	79.64%	63.26%
3D ResNext [3]	55.14%	80.36%	67.75%
I3D [1]	58.37%	83.74%	71.06%
TSM [18]	61.55%	85.05%	73.30%
R(2+1)D-34 [14]	62.94%	85.71%	74.33%
I3D+SE [12]	64.38%	86.55%	64.23%
I3D+NL [9]	65.82%	87.49%	76.66%
ip-CSN-152 [16]	65.90%	87.02%	76.46%
3D ResNext+SCA(Ours)	64.90%	83.02%	73.96%
I3D+SCA(Ours)	**67.22%**	**88.28%**	**77.75%**

Table 2:
Comparison with the State-of-the-Art methods on UCF-101.