SCA Net: Sparse Channel Attention Module for Action Recognition

Hang Song*, YongHong Song, YuanLin Zhang

College of Artificial Intelligence, Xi'an Jiaotong University

Background

Action recognition aims at determining the category of the action in a trimmed video, which belongs to the classification problem defined by traditional machine learning.

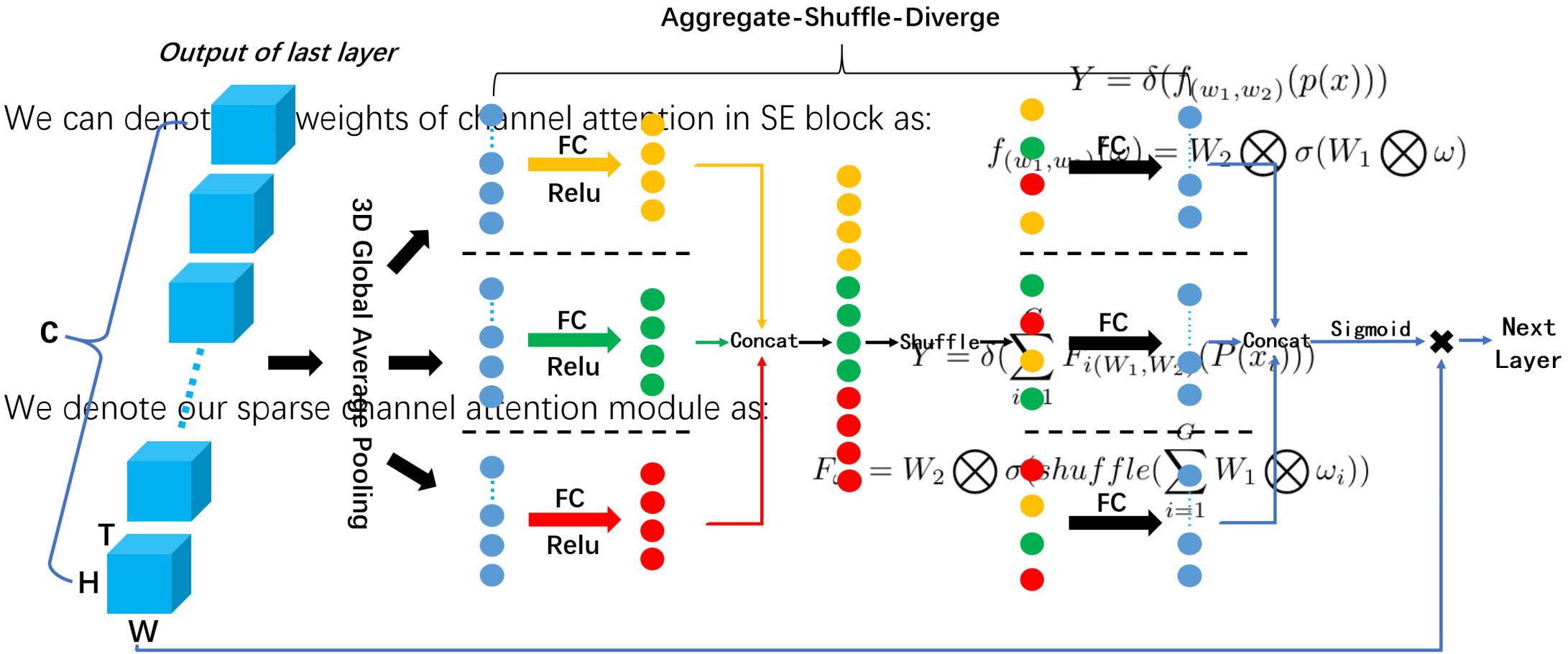
Motivation

- Recently, channel attention method has arose researchers' interests, as it brings a great performance gain when it is incorporated into all kinds of convolution blocks.

- However, in the video domain, the research on the channel attention method hasn't attracted much attention, which is mainly due to the high computational burden of video analysis.

- In order to make good use of channel attention in the video domain, we must solve the following problem: How to learn channel attention with a more lightweight method?

Method



We can denot weights of channel attention in SE block as:

$$Y = \delta(f_{(w_1,w_2)}(p(x)))$$

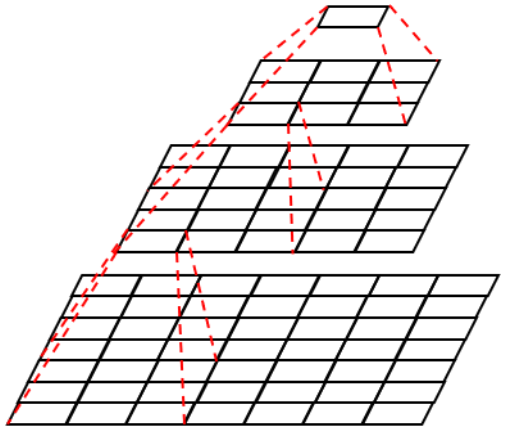
$$f_{(w_1,w_2)}(\omega) = W_2 \bigotimes \sigma(W_1 \bigotimes \omega)$$

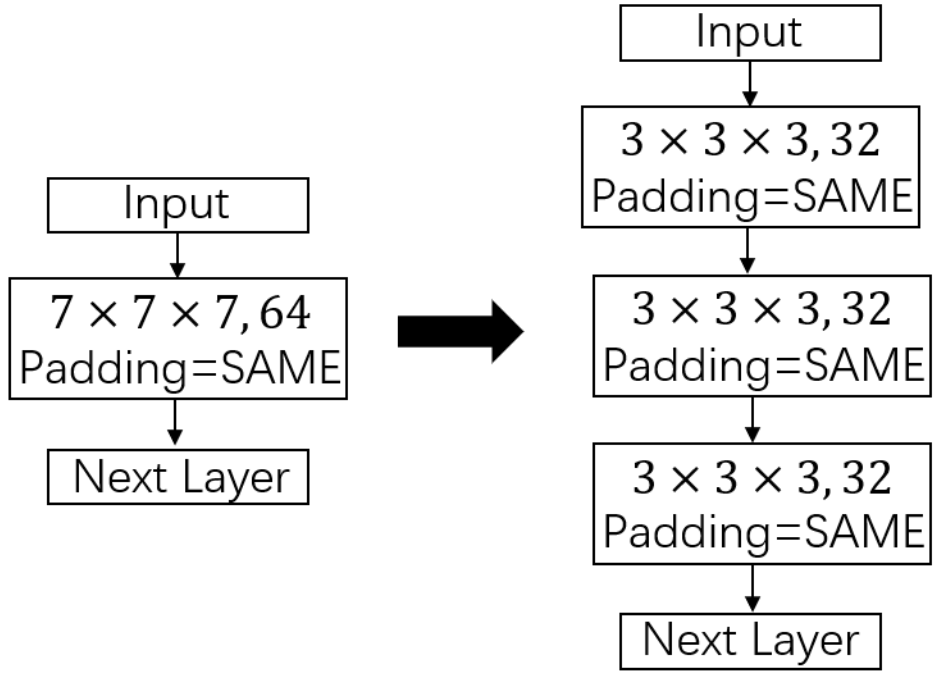
We denote our sparse channel attention module as:

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

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

Method




Input
7 × 7 × 7, 64
Padding=SAME
Next Layer
Input
3 × 3 × 3, 32
Padding=SAME
3 × 3 × 3, 32
Padding=SAME
3 × 3 × 3, 32
Padding=SAME
Next Layer

Datasets

- **UCF-101** is an action recognition dataset of realistic action videos, collected from YouTube, having **101** action categories with **13320** videos. The videos in 101 action categories are grouped into 25 groups, where each group can consist of 4-7 videos of an action. The action categories include several types such as: human-object interaction, body-motion only, human-human interaction, playing musical instruments, sports, etc.

- **HMDB-51** contains **51** action categories with **6849** clips and each containing a minimum of 101 clips, collected from various sources, mostly from movies, and a small proportion from public databases such as the Prelinger archive, YouTube and Google videos.

Experiment Results

西安交通大學 XI'AN JIAOTONG UNIVERSITY

TABLE I: Top-1 accuracy on UCF-101/HMDB-51 and parameter size for SCA module at different group numbers with $r = 2$.

Methods	Para.	Para.*	UCF-101	HMDB-51
Factorized I3D	0	0	60.23%	32.48%
+SCA-G_2	$C^2/2$	1.20M	**67.22%**	**36.52%**
+SCA-G_4	$C^2/4$	0.60M	66.40%	35.18%
+SCA-G_8	$C^2/8$	0.30M	65.35%	34.59%
+SCA-G_{16}	$\boldsymbol{C^2/16}$	**0.15M**	64.10%	33.76%

TABLE II: Top-1 accuracy on UCF-101/HMDB-51 and parameter size for SCA module at different ratios with $G = 2$.

Methods	Para.	Para.*	UCF-101	HMDB-51
Factorized I3D	0	0	60.23%	32.48%
+SCA-r_2	$C^2/2$	1.20M	**67.22%**	**36.52%**
+SCA-r_4	$C^2/4$	0.60M	66.83%	35.24%
+SCA-r_8	$C^2/8$	0.30M	64.97%	34.16%
+SCA-r_{16}	$\boldsymbol{C^2/16}$	**0.15M**	63.82%	33.24%

TABLE III: SCA module with/without ASD function

Methods	Para.	Para.*	UCF-101	HMDB-51
Factorized I3D	0	0	60.23%	32.48%
Without ASD	$C^2/4$	0.60M	63.27%	33.03%
With ASD	$C^2/4$	0.60M	**66.40%**	**35.18%**

TABLE V: Effects of 3D Convolution Factorization on UCF-101 and HMDB-51.

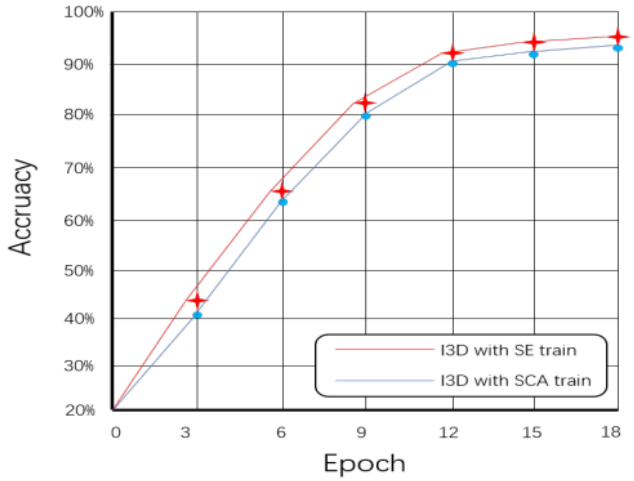
Methods	Para.	UCF-101	HMDB-51
I3D	343	58.37%	31.24%
Factorized I3D	**81**	**60.23%**	**32.48%**
3D ResNext	343	55.14%	30.52%
Factorized 3D ResNext	**81**	**57.36%**	**31.48%**

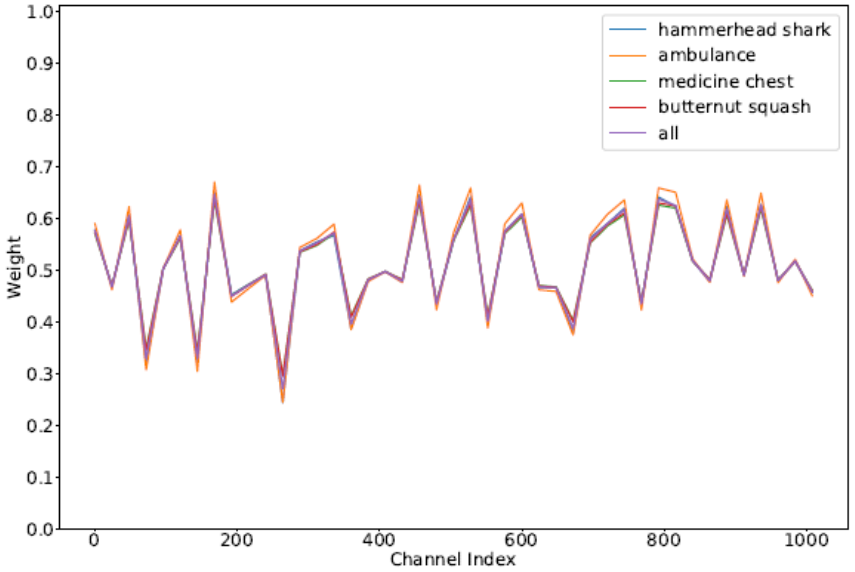
Experiment Results

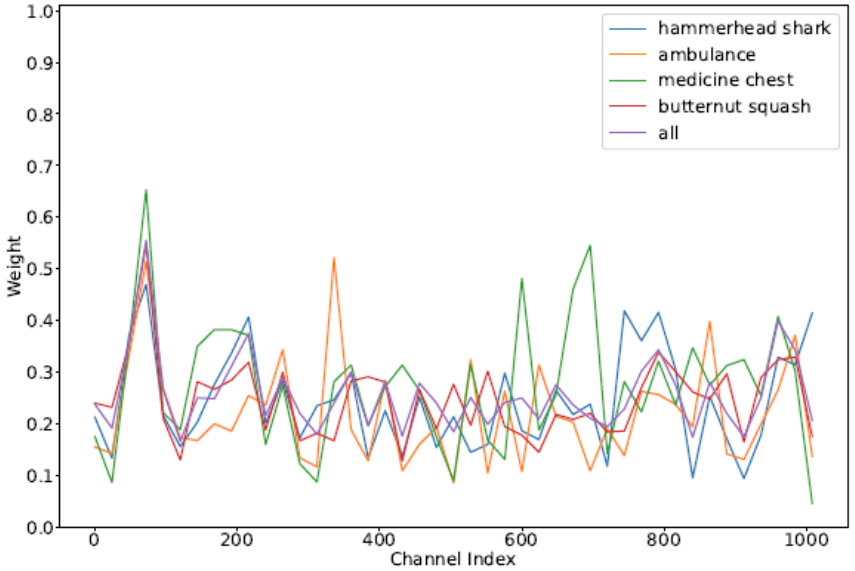
西安交通大學 XI'AN JIAOTONG UNIVERSITY

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

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%







Experiment Results

TABLE VI: Comparison with the State-of-the-Art methods on UCF-101 split1 training from scratch.

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%**

Conclusion

We aim to reduce complexity of traditional channel attention methods and widely used mainstream 3D CNNs. To this end, a novel Sparse Channel Attention (SCA) module has been presented, which adopts the idea of sparse channel connection to generate channel attention weights. In addition, we design an Aggregate-Shuffle-Diverge function to enhance cross-group interaction.

Meanwhile, we also employ 3D convolution factorization in mainstream 3D CNNs to further reduce parameters. The experimental results demonstrate that our SCA module is a more efficient and effective method compared with traditional channel attention methods, and it achieves a significant performance gain when it is incorporated into various backbone networks.

In future, we will further investigate our SCA module with 2D deep CNN architectures and demonstrate the correlations with other channel attention methods.