ULSAM: Ultra-Lightweight Subspace Attention Module for Compact Convolutional Neural Networks

Rajat Saini[1*], Nandan Kumar Jha[1*], Bedanta Das[1], Sparsh Mittal[2], C. Krishna Mohan[1]

[1]Indian Institute of Technology Hyderabad,

[2]Indian Institute of Technology Roorkee (*equal contribution)

Introduction and Motivation

- The **locality** of convolution in deep networks offers a theoretical guarantee to avoid the *curse of dimensionality* for approximating the hierarchically local compositional functions.
- To capture **global dependencies** networks are made deeper (that enlarge the effective receptive field size) and incur higher **#FLOPs** and **#parameters**.
- Convolution is a **linear** operator and to capture non-linear abstractions in input CNNs employed large number of filters which increases computational complexity and parameter overhead.
- Self-attention mechanism offers *infinite receptive field size* and captures global dependencies in a **compute efficient manner** hence remove the inefficiencies of convolution operation.

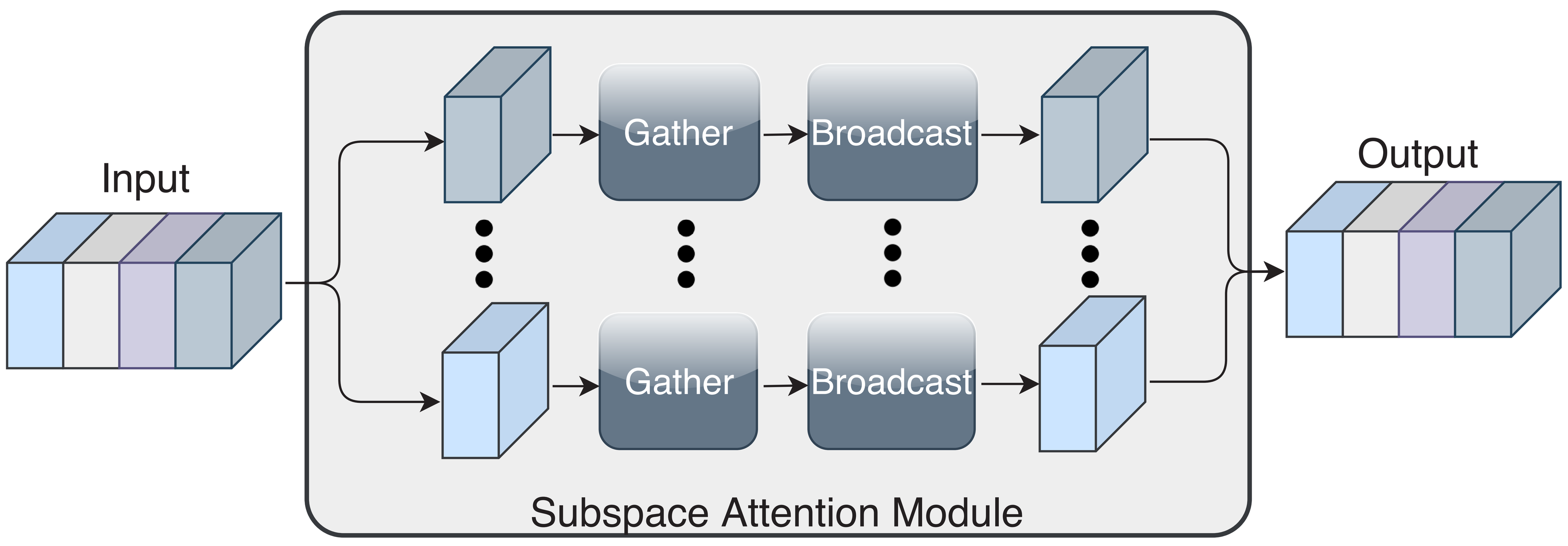
Challenges

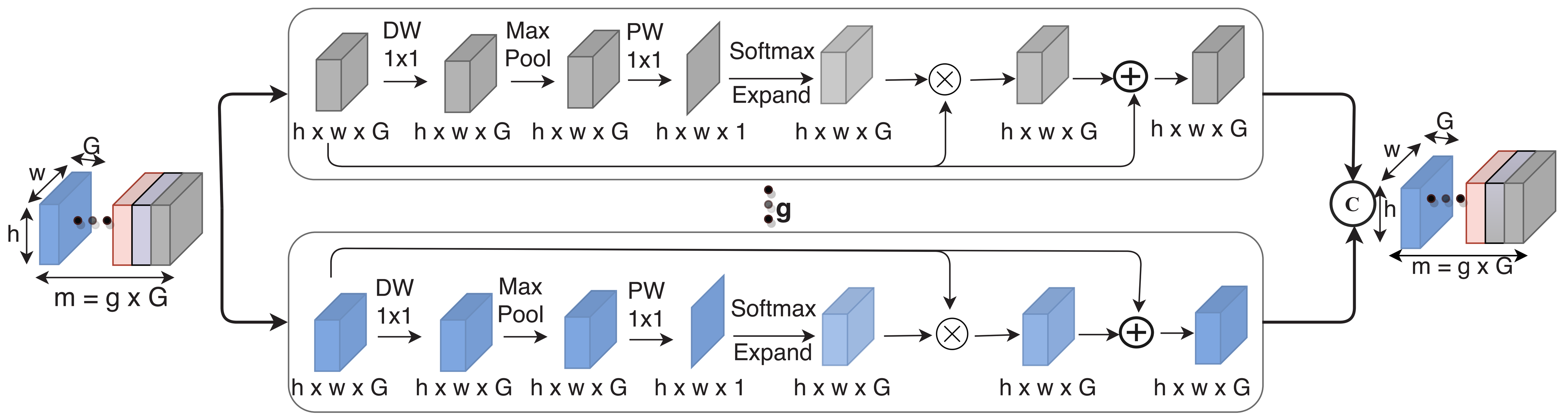
- State-of-the-art attention mechanisms incur **higher compute overhead** due to 1×1 conv (for generating attention maps) and/or **parameter overhead** due to the use of MLP (which learns the cross-channel interaction) and **undesirable** for compact CNNs such as MobileNet-V1/V2.

(Comparison with $m = 512$, $t = \frac{m}{8}$, $r = 16$, $h \times w = 14 \times 14$, and dilation rate is 4 in BAM)

Attention module	MLP	1 × 1 conv	#Params	#FLOPs	#Params ($\times 10^3$)	#FLOPs ($\times 10^6$)
Non-local [Wang, CVPR'18]	×	✓	$2m^2$	$2m^2hw$	524	102.76
A^2 - Net [Chen, NeurIPS'18]	×	✓	$2mt$	$2mthw$	66	12.85
SE-Net [Hu, CVPR'18]	✓	×	$\frac{2m^2}{r}$	$\frac{2m^2}{r}$	33	0.03
BAM [Park, BMVC'18]	✓	✓	$\frac{4m^2}{r} + \frac{18m^2}{r^2}$	$\frac{2m^2}{r} + (\frac{4m^2}{r} + \frac{18m^2}{r^2})hw$	84	16.49
CBAM [Woo, ECCV'18]	✓	×	$\frac{2m^2}{r} + 98$	$\frac{2m^2}{r} + 98hw$	33	0.05

Proposed Method: Subspace Attention Module





Salient features of ULSAM

- ULSAM exploits the *linear relationship* between feature map subspace and avoids the use of **parameter-heavy** MLP.
- Generating separate attention maps for different parts of feature map space enable **multi-scale** (desirable for object detections with objects of different scale) and **multi-frequency** (desirable for fine-grained image classification tasks) feature representation.

Attention module	subspace attention	MLP	1 × 1 conv	#Params ($\times 10^3$)	#FLOPs ($\times 10^6$)	#Params (*norm.*)	#FLOPs (*norm.*)
Non-local [Wang, CVPR'18]	×	×	✓	524	102.76	512×	512×
A^2 - Net[Chen, NeurIPS'18]	×	×	✓	66	12.85	64×	64×
SE-Net [Hu, CVPR'18]	×	✓	×	33	0.03	33×	0.16×
BAM [Park, BMVC'18]	×	✓	✓	84	16.49	82×	82.16×
CBAM [Woo, ECCV'18]	×	✓	×	33	0.05	33×	0.26×
ULSAM (**ours**)	✓	×	×	1	0.2	1×	1×

ULSAM reduces both the **computational complexity** and the **number of parameters** and hence suitable for deployment in compact CNNs.

Results on ImageNet1k

Model	**#Params**	**#FLOPs**	$g = 1(\%)$	$g = 2(\%)$	$g = 4(\%)$	$g = 8(\%)$	$g = 16(\%)$
1.0 MV1 (vanilla)	4.2M	569M	Top-1 = 70.65				
1.0 MV1 + ULSAM	4.2M	569.2M	70.69	70.84	70.77	70.59	**70.89**
1.0 MV1 + ULSAM	4.2M	569.2M	70.62	70.88	70.61	**70.92**	70.73
1.0 MV1 + ULSAM	4.2M	569.1M	70.63	70.85	**70.86**	70.74	70.82
MV2 (vanilla)	3.4M	300M	Top-1 = 71.25				
MV2 + ULSAM	3.4M	300.01M	71.31	71.39	**71.64**	71.35	71.42

@g = 4, top-1 accuracy of MV1(MV2) increased by **0.27% (0.39%)** with negligible compute overhead.

Fine-grained classification

MobileNet-V1 on Food-101, Birds, and Dogs dataset

Models	**#Params**	**#FLOPs**	**Food-101**	**Birds**	**Dogs**
MV1 (vanilla)	4.2M	569M	81.31	62.88	62.20
MV1 + ULSAM ($g = 1$)	3.9M	517M	81.28	62.46	62.73
MV1 + ULSAM ($g = 4$)	3.9M	517M	81.30	63.52	63.06
MV1 + ULSAM ($g = 8$)	3.9M	517M	81.19	**64.44**	**63.30**
MV1 + ULSAM ($g = 16$)	3.9M	517M	**81.62**	63.47	62.75

MobileNet-V2 on Birds dataset

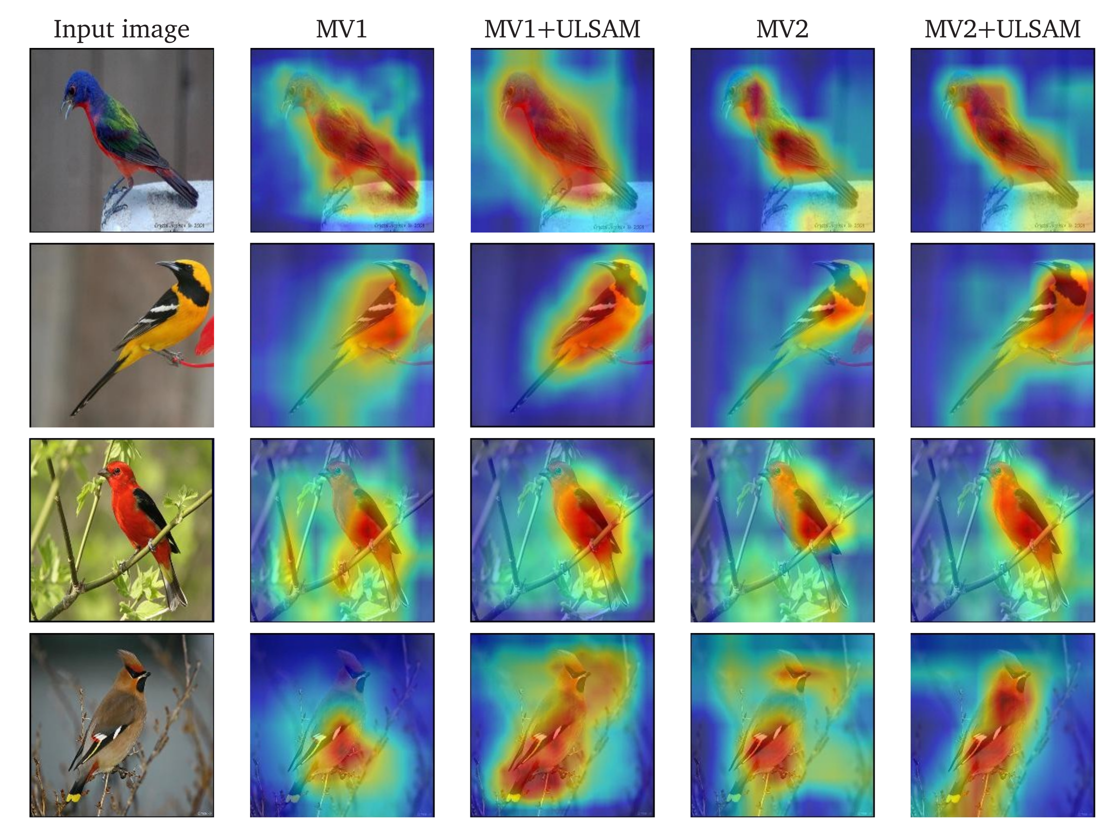
Model	**#Params**	**#FLOPs**	$g = 1$	$g = 4$	$g = 8$	$g = 16$
MV2 (vanilla)	3.4M	300M	Top-1 = 62.94			
MV2 + ULSAM	3.28M	277.34M	63.01	63.05	63.11	**64.32**
MV2 + ULSAM	3.08M	284.54M	63.98	64.44	**65.03**	63.47
MV2 + ULSAM	3.23M	267.06M	63.43	**63.47**	63.10	62.21
MV2 + ULSAM	2.77M	269.08M	64.19	64.57	64.61	**65.03**
MV2 + ULSAM	2.97M	261.88M	63.35	64.70	**65.41**	63.31
MV2 + ULSAM	2.54M	224.16M	64.11	**64.15**	63.22	63.98

Ablation Study on ImageNet1k

Models	**#Params**	**#FLOPs**	**Top-1**	**Top-5**
1.0 MV1 (vanilla)	4.2M	569M	70.65	89.76
1.0 MV1 + ULSAM ($g = 1$)	3.9M	517M	69.92	89.25
1.0 MV1 + ULSAM ($g = 2$)	3.9M	517M	70.14	89.67
1.0 MV1 + ULSAM ($g = 4$)	3.9M	517M	**70.43**	89.92
1.0 MV1 + ULSAM ($g = 8$)	3.9M	517M	70.29	89.96
1.0 MV1 + ULSAM ($g = 16$)	3.9M	517M	70.04	**89.98**
0.75 MV1 (vanilla)	2.6M	325M	67.48	88.00
0.75 MV1 + ULSAM ($g = 1$)	2.4M	296M	**67.98**	88.06
0.75 MV1 + ULSAM ($g = 4$)	2.4M	296M	67.81	**88.43**
0.50 MV1 (vanilla)	1.3M	149M	63.22	84.63
0.50 MV1 + ULSAM ($g = 1$)	1.2M	136M	**63.42**	84.70
0.50 MV1 + ULSAM ($g = 4$)	1.2M	136M	63.25	**84.81**
MV2 (Vanilla)	3.4M	300M	71.25	90.19
MV2 + ULSAM ($g = 4$)	2.96M	261.88M	**71.52**	**90.25**
MV2 + ULSAM($g = 4$)	2.77M	269.07M	70.74	89.15
MV2 + ULSAM ($g = 4$)	2.54M	223.77M	69.72	87.79

MV2 achieves a **25% (13%)** reduction in #parameters (#FLOPs) with **0.27%** improvement in top-1 accuracy.

Attention Visualization



Conclusion

- Single attention map for entire feature space **does not capture** the interaction between the different feature map subspace.
- Optimum number of attention maps are required to **maximize** the predictive performance of networks.

Contact

cs17mtech11010 [at] iith [dot] ac [dot] in