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. K. Mohan[1]

[1]{cs17mtech11002, cs17mtech11010, cs17mtech11009, ckm}@iith.ac.in,
[2]sparshfec@iitr.ac.in (*equal contribution)

[1]Indian Institute of Technology Hyderabad
[2]Indian Institute of Technology Roorkee

IEEE Winter Conference on Applications of Computer Vision 2020

March 3, 2020

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

WACV 2020

Convolution: The winter of despair

Convolution is a **linear** operator and captures **local dependencies** in feature space.

- Inefficiencies of convolution operation
  - Limited receptive field size $\Longrightarrow$ Deeper networks to capture long range dependencies.
  - Captures linear abstraction $\Longrightarrow$ Wider networks to capture non-linear abstractions in the input data.

Convolution: The winter of despair

Convolution is a **linear** operator and captures **local dependencies** in feature space.

- Inefficiencies of convolution operation
  - Limited receptive field size $\Longrightarrow$ Deeper networks to capture long range dependencies.
  - Captures linear abstraction $\Longrightarrow$ Wider networks to capture non-linear abstractions in the input data.

Challenges: Deeper and wider networks lead to **higher** computational complexity, memory footprint, and energy consumption; **inefficient** back-propagation; **increased** serialization and **lack** of parallelizability.

Solution?

Solution?

Self-attention: A spring of hope

Self-attention mechanism in computer vision models offers infinite receptive field size and captures global dependencies in feature space.

Key advantage: Employing attention mechanism in deeper layers of CNNs **enlarge** the effective receptive field size and enable compute and parameter **efficient** feature representation.

SOTA self-attention in computer vision models

Table: Compute and parameter overheads of different attention modules

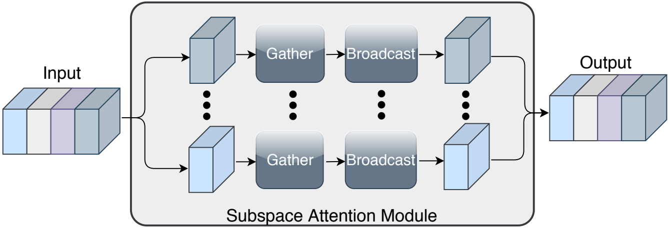
Attention module	MLP	1×1 conv	#Params	#FLOPs	#Params ($\times 10^3$)	#FLOPs ($\times 10^6$)
Non-local [Wang, CVPR'18]	$\times$	$\checkmark$	$2m^2$	$2m^2hw$	524	102.76
A^2 - Net [Chen, NeurIPS'18]	$\times$	$\checkmark$	$2mt$	$2mthw$	66	12.85
SE-Net [Hu, CVPR'18]	$\checkmark$	$\times$	$\frac{2m^2}{r}$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

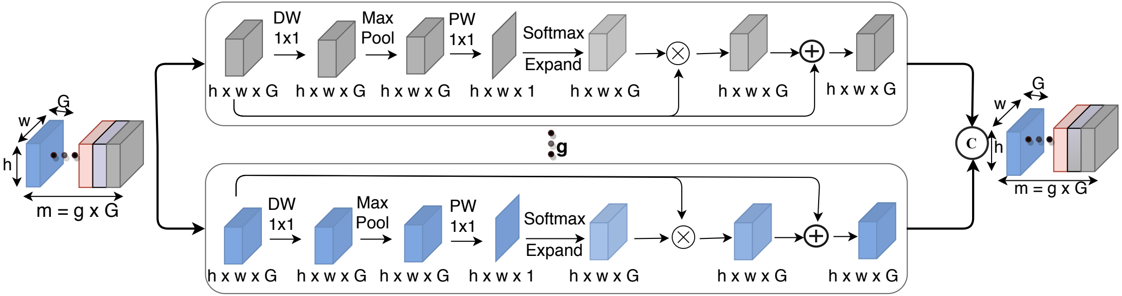
SOTA attention mechanism uses:

- 1×1 convolution: To generate attention maps.
- MLP: To model the cross-channel dependencies.

Key observation: SOTA attention mechanism suitable for large and over-parameterized CNNs and **undesirable** for compact CNNs.

Proposed Mechanism: Subspace attention mechanism





Salient features of ULSAM

- Exploits the **linear relationship** between feature map subspace.
  - No need of parameter-heavy MLP.
- Enables **multi-scale** feature representation.
  - Desirable when objects of different sizes in frame.
- Enables **multi-frequency** feature representation.
  - Desirable when discriminative regions contain high frequency features.

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.

Experimental Results: MobileNetV1/V2 on ImageNet-1K

Model	#Params	#FLOPs	$g=1$ Top-1	$g=1$ Top-5	$g=2$ Top-1	$g=2$ Top-5	$g=4$ Top-1	$g=4$ Top-5	$g=8$ Top-1	$g=8$ Top-5	$g=16$ Top-1	$g=16$ Top-5
1.0 MV1 (vanilla)	4.2M	569M	Top-1 = 70.65, Top-5 = 89.76									
1.0 MV1 + ULSAM	4.2M	569.2M	70.69	89.85	70.84	89.87	70.77	**89.91**	70.59	89.83	**70.89**	89.74
1.0 MV1 + ULSAM	4.2M	569.2M	70.62	89.86	70.88	89.88	70.61	89.79	**70.92**	**89.98**	70.73	89.78
1.0 MV1 + ULSAM	4.2M	569.1M	70.63	89.60	70.85	89.97	**70.86**	89.85	70.74	89.81	70.82	**90.05**
MV2 (vanilla)	3.4M	300M	Top-1 = 71.25, Top-5 = 90.19									
MV2 + ULSAM	3.4M	300.01M	71.31	90.28	71.39	90.34	**71.64**	90.27	71.35	90.36	71.42	**90.43**

Key observations:

- A significant gain in the accuracy of MV1/MV2 when $g > 1$.
- Accuracy of MV1/MV2 is higher (compared to baseline model) when $g \geq 4$.
- @ $g = 4$, top-1 accuracy of MV1(MV2) increased by 0.27% (0.39%)

Key Takeaway: Separate attention maps for the different parts (subspace) of feature maps helps in better feature representation

MobileNetV1/V2 on fine-grained image classification datasets

Models	#Params	#FLOPs	Food-101		Birds		Dogs	
			Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
MV1 (vanilla)	4.2M	569M	81.31	95.24	62.88	86.05	62.20	89.66
MV1 + ULSAM ($g = 1$)	3.9M	517M	81.28	**95.50**	62.46	86.01	62.73	88.80
MV1 + ULSAM ($g = 4$)	3.9M	517M	81.30	95.37	63.52	85.80	63.06	89.58
MV1 + ULSAM ($g = 8$)	3.9M	517M	81.19	95.41	**64.44**	**86.60**	**63.30**	**89.68**
MV1 + ULSAM ($g = 16$)	3.9M	517M	**81.62**	95.33	63.47	84.90	62.75	89.35

Model	#Params	#FLOPs	$g = 1$		$g = 4$		$g = 8$		$g = 16$	
			Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
MV2 (vanilla)	3.4M	300M	Top-1 = 81.51, Top-5 = 95.24							
MV2 + ULSAM	3.28M	277.34M	81.67	**95.82**	81.71	95.47	81.76	95.51	**81.94**	95.63
MV2 + ULSAM	3.08M	284.54M	82.05	**95.56**	**82.02**	95.48	81.74	95.40	81.54	95.14
MV2 + ULSAM	2.97M	261.88M	81.57	**95.44**	81.69	95.36	**82.13**	95.42	81.84	95.40
MV2 + ULSAM	2.54M	224.16M	82.38	95.76	82.31	95.80	82.59	**95.82**	**82.91**	95.77

Key observation: The accuracy of both MV1 and MV2 is higher than the baseline model with significantly lower computation and number of parameters.

Ablation study: MV1 and MV2 on ImageNet-1K

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**

Models	#Params	#FLOPs	Top-1	Top-5
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

Key observations:

- MV2 achieves a **25% (13%)** reduction in number of parameters (computation) with **0.27%** improvement in top-1 accuracy.
- 0.75-MV1 (0.50-MV1) achieves a **9.1% (8.9%)** reduction in computational complexity with **0.5% (0.1%)** improvement in top-1 accuracy.

Conclusion

- One single attention map in entire feature space does not capture the subspace relationship between the different feature subspace.
- Subspace attention module is a more efficient way to learn the cross-channel interaction in feature maps space of networks.
- Optimum number of attention maps are required to maximize the predictive performance of networks.
- Learning separate attention maps for different feature subspace enables multi-scale and multi-frequency feature representation.
- Multi-frequency feature representation is more desirable for fine-grained image classification tasks.

Thanks for your attention...!!!