SEP-Nets: Small and Effective Pattern Networks

Zhe Li[1,2], Xiaoyu Wang[1], Xutao Lv[1], Tianbao Yang[2]

Snap Research[1] The University of Iowa[2]

Highlights

- Proposed **parttern binarization** method.
- Designed a new **pattern residual block**.
- Designed a novel **SEP-Net Module**.
- Proposed **S**mall and **E**ffective **P**attern **N**etworks.
- Achieved the-state-of-art performance.

Pattern Binarization

- $k \times k$ ($k > 1$) filters serve as spatial pattern extraction.
- 1×1 filters serve as data transformation.
- Reduced number of parameters in model dramatically.

Averaged Quantization Error of Different Sized Filters from GoogLeNet.

1×1	3×3	5×5
0.0462	0.0029	0.0056

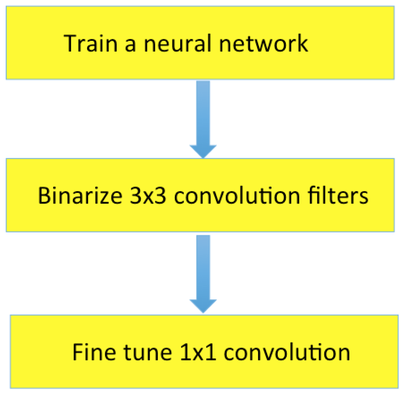
-0.0219	0.0408	-0.0547
-0.0855	0.0478	-0.0510
-0.0105	0.0924	-0.0126

→

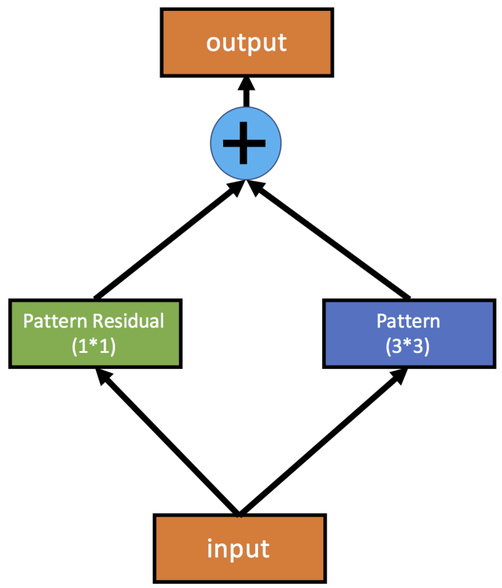
-1	1	-1
-1	1	-1
-1	1	-1

A trained 3×3 filter from GoogLeNet (Left), and its binarized version (Right)

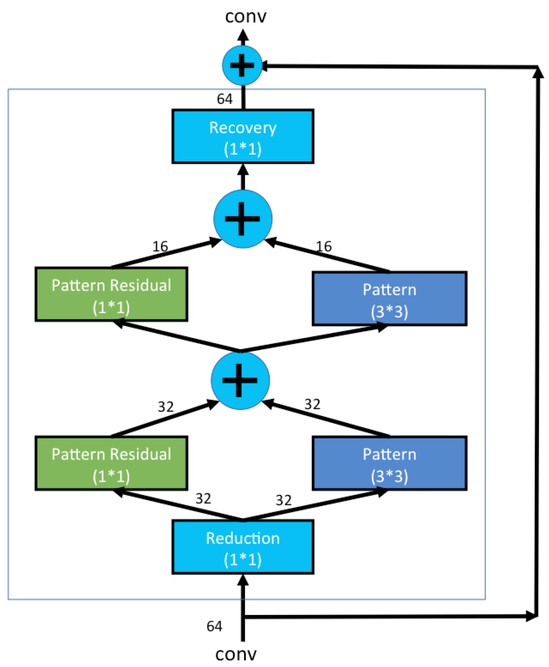
Remark: can be easily adopted to any successful networks structures such as GoogleNet, ResNet inlcuding the designed **SEP-Nets** as following procedure:



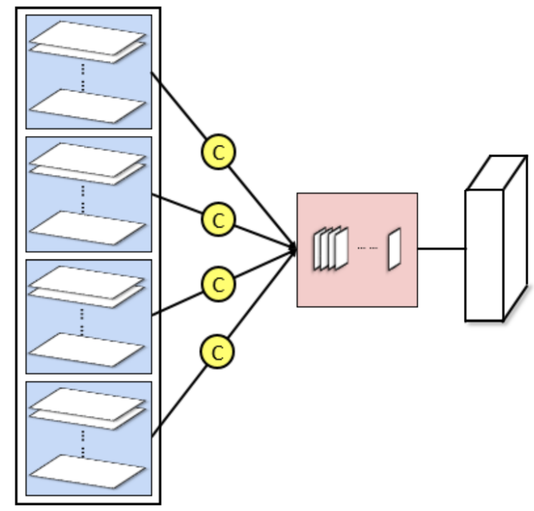
Ingredients for SEP-Nets



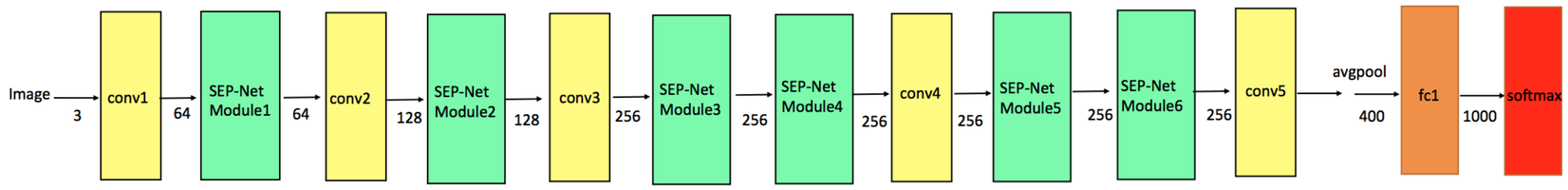
Pattern Residual Block



SEP-Net Module



Group-wise convolution



The architecture of our experimented SEP-Nets

Experimental Results

model	Parameter number†	Top1-Top5†	Parameter number‡	Top1-Top5‡
GoolgeNet	6.99M	0.6865 0.8891	4.43M	0.6797 0.8827
Customized-Inception	5.10M	0.6480 0.8630	2.43M	0.6400 0.8550

†: full model

‡: model after binarizing 3×3 convolutional filters and fine tuning

SEP-Net-R: SEP-Net with raw valued weights

SEP-Net-B: SEP-Net with pattern binarization

SEP-Net-BQ: SEP-Net with pattern binarization and other weights quantized using linear quantization with 8 bits

Model	Parameter number	Size (bytes)	Top-1 Acc
MobileNet	1.3M	5.2MB	0.637
	2.6M	10.4MB	0.684
SEP-Net-R	1.3M (small)	5.2MB	0.658
	1.7M (large)	6.7MB	0.667
SqueezeNet	1.2M	4.8MB	0.604
MobileNet	1.3M	5.2MB	0.637
SEP-Net-R (Small)	**1.3M**	**5.2MB**	**0.658**
SEP-Net-B (Small)	1.1M	4.2MB	0.637
SEP-Net-BQ (Small)	1.1M	1.3MB	0.635