ARCHITECTURE-AWARE NETWORK PRUNING FOR VISION QUALITY APPLICATIONS

Wei-Ting Wang, Han-Lin Li, Wei-Shiang Lin, Cheng-Ming Chiang, Yi-Min Tsai

MediaTek Inc.

{wei-ting.wang, Han-Lin.Li, Wei-Shiang.Lin, Jimmy.Chiang, Yi-Min.Tsai}@mediatek.com

MediaTek AI

Paper

INTRODUCTION

- **Goal**
  - Deploy neural networks on resource-constrained systems for vision quality applications without special hardware design
  - Minimize Multiply-Accumulate (MAC) and memory bandwidth (BW) without quality-metric drops
- **Challenge of Vision Quality Applications**
  - Increased MAC per inference because of large resolution

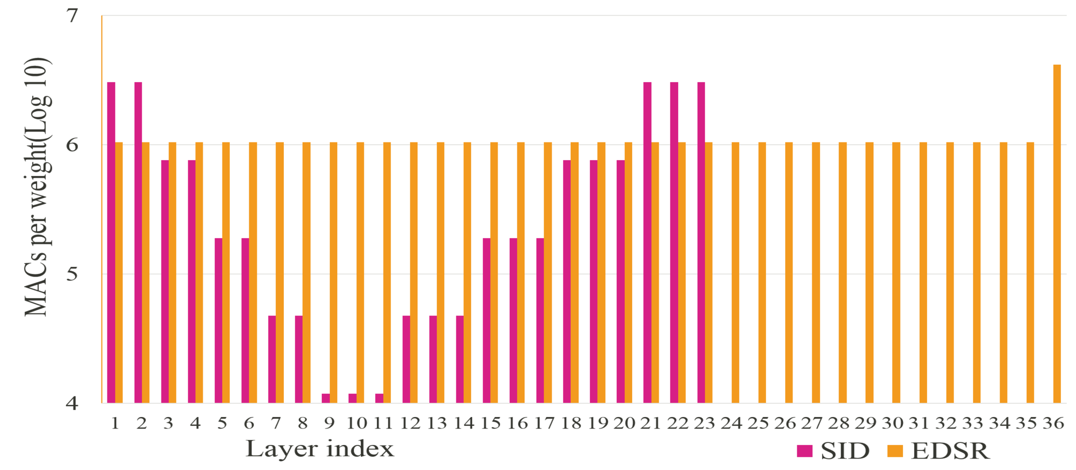
Application	Network	Resolution	# of GMAC
Classification	MobileNet-v1	224x224	0.6
Low-light photography	SID [1]	1424x2128	1500
Super-resolution	EDSR [2]	1020x1020	560

[1] Learning to see in the dark.

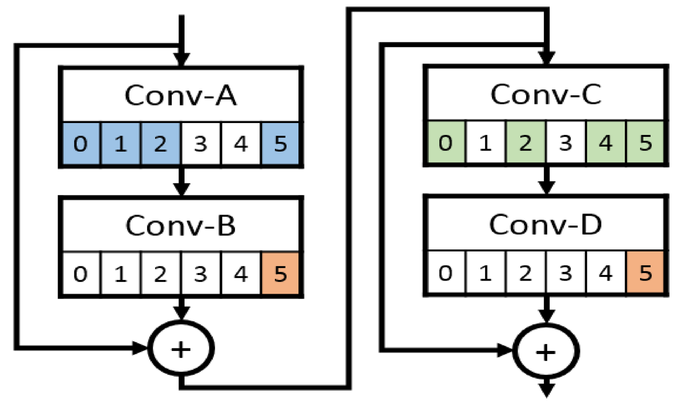
[2] Enhanced deep residual networks for single image super-resolution. Baseline (Single-scale) x2

- **Solution**
  - Learning network architecture with performance target by adaptive pruning threshold while keeping quality

NETWORK ARCHITECTURE ANALYSIS



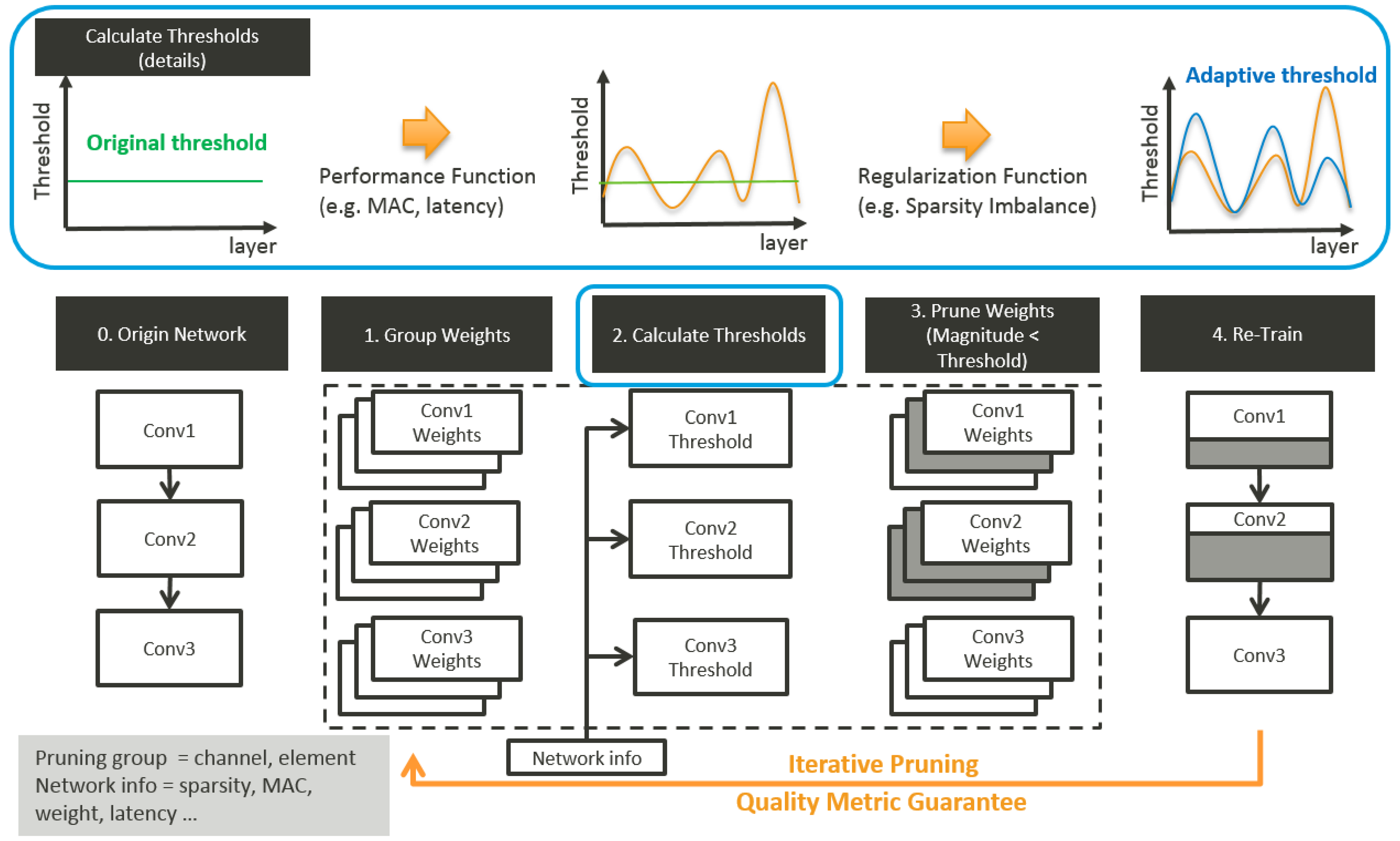
SID:
MAC/weight are much larger on both top and bottom layers



EDSR:
Output channel of a given layer (Conv-D) and its preceding layer (Conv-B) should be grouped and pruned on residual blocks

ARCHITECTURE-AWARE NETWORK PRUNING

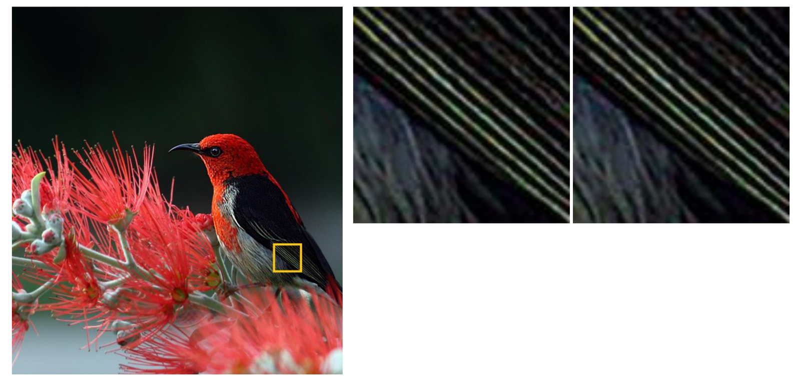
Observation	Method
MAC is 5 order of magnitude larger than weight	Enhance MAC Efficiency
A layer removed can severely degrade the quality	Keep Layer Depth
A residual block (16-layer-group, 43% MAC) are hard to be pruned	Balance Pruned Output Channel



PRUNED SUBJECTIVE QUALITY

SID:
Original (Left), Pruned (Right)

EDSR:
Original (Left,Mid), Pruned (Right)



PRUNED COMPLEXITY AND QUALITY-METRIC

- MAC of SID and EDSR are reduced by 58% and 37%
- BW of convolutional layer are reduced by 20% to 40%
- Without degradation of PSNR, SSIM and subjective quality

Method	Description
A	Iterative Prune
B	A + Keep Layer Depth
C	B + Enhance MAC Efficiency
D	C + Balance Pruned Output Channel

Table 1. Detailed results. BW, considering only convolutional layers, consists of both weights and activations. Each weight and activation is represented with 4-byte floating-point numerical precision.

Network	Solution	# of MAC ($\times 10^9$)		# of Weights ($\times 10^3$)		# of Activations ($\times 10^6$)		BW (MByte/Inference)		Validation PSNR	Validation SSIM
SID	Original	560	(100%)	7757	(100%)	1915	(100%)	1922	(100%)	28.54	0.767
SID	Method-A	458	(82%)	6918	(89%)	1632	(85%)	1639	(85%)	28.54	0.768
SID	Method-B	354	(63%)	5275	(68%)	1485	(78%)	1491	(78%)	28.54	0.771
SID	Method-C	270	(48%)	5584	(72%)	1219	(64%)	1225	(64%)	28.54	0.769
SID	Method-D	236	(42%)	4241	(55%)	1169	(61%)	1173	(61%)	28.55	0.768
EDSR	Original	1428	(100%)	1367	(100%)	5076	(100%)	5077	(100%)	34.42	0.942
EDSR	Method-A	1085	(76%)	1037	(76%)	4481	(88%)	4481	(88%)	34.43	0.942
EDSR	Method-B	1085	(76%)	1037	(76%)	4481	(88%)	4481	(88%)	34.43	0.942
EDSR	Method-C	1085	(76%)	1037	(76%)	4481	(88%)	4481	(88%)	34.43	0.942
EDSR	Method-D	897	(63%)	857	(63%)	4083	(80%)	4083	(80%)	34.42	0.942

PRUNED NETWORK ARCHITECTURE ANALYSIS

- Pruned output channel per layer: SID (top), EDSR (bottom)