IMPQ: Reduced Complexity Neural Networks via Granular Precision Assignment

Sujan Kumar Gonugondla*

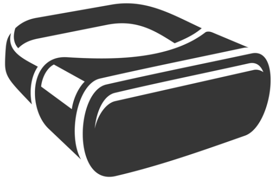
Naresh Shanbhag

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Urbana, IL-61801
gonugon2@illinois.edu
*Now with Amazon Web Services

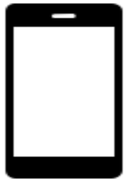
Machine Learning Under Resource Constraints

smart cities

self-driving cars

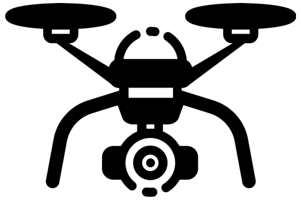


AR/VR

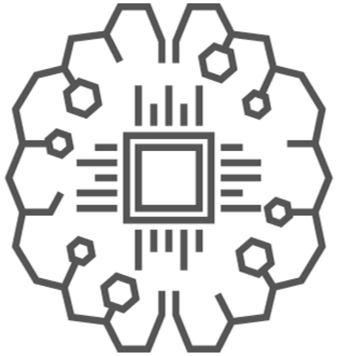


smart devices

wearables



autonomous systems



smart homes

decision making under *resource constraints*

limited energy supply, storage, real-time response

Machine learning at the edge opens up many interesting applications

Approach to improve inference efficiency

Novel hardware accelerators

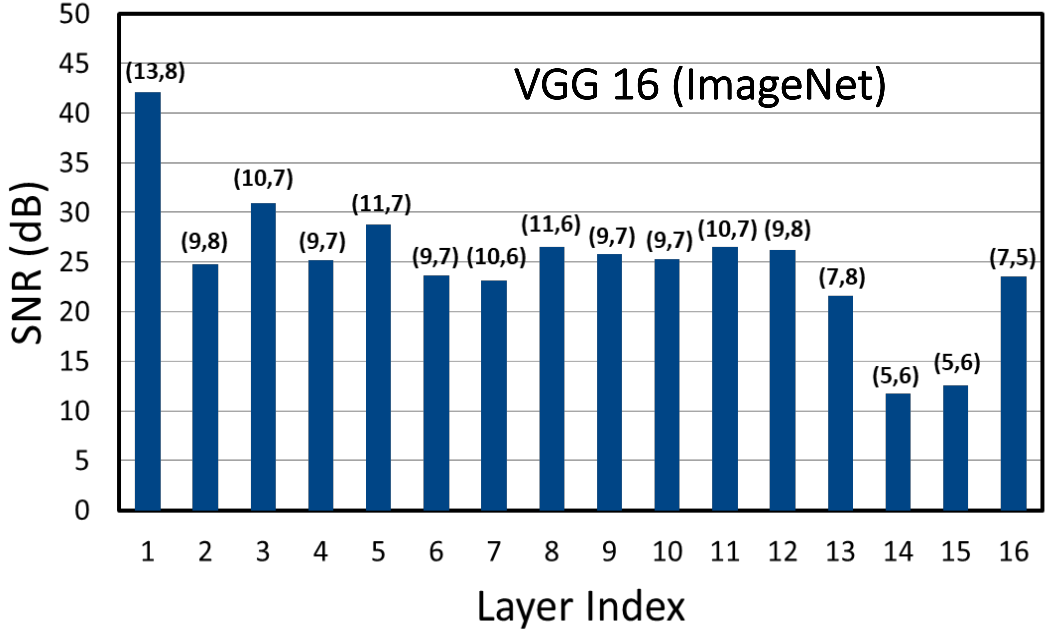
- specialized neural-network accelerators, and drivers
- in-memory computing

Novel algorithmic approaches

- Efficient neural network architectures
  - Ex: using depth-wise separable layers, low-rank approximations
- Knowledge distillation
- Pruning
- Low-precision quantization

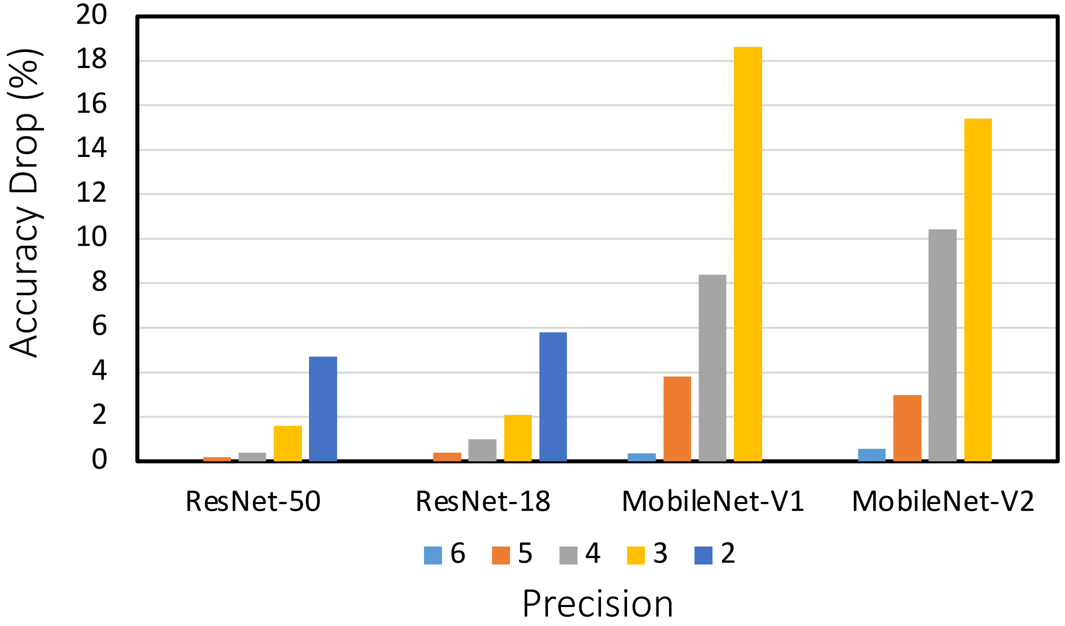
Precision/SNR Requirements in Neural Nets

pretrained floating point network →
< 1% loss in accuracy [Sakr *et al.*, ICML-17]



[Wang *et al.*, CVPR-19] & [Choi *et al.*, arxiv-18]

#operations	3.8G	1.7G	575M	300M
#parameters	23M	11M	4.2M	3.4M
FL accuracy	76.9	70.2	70.82	71.81

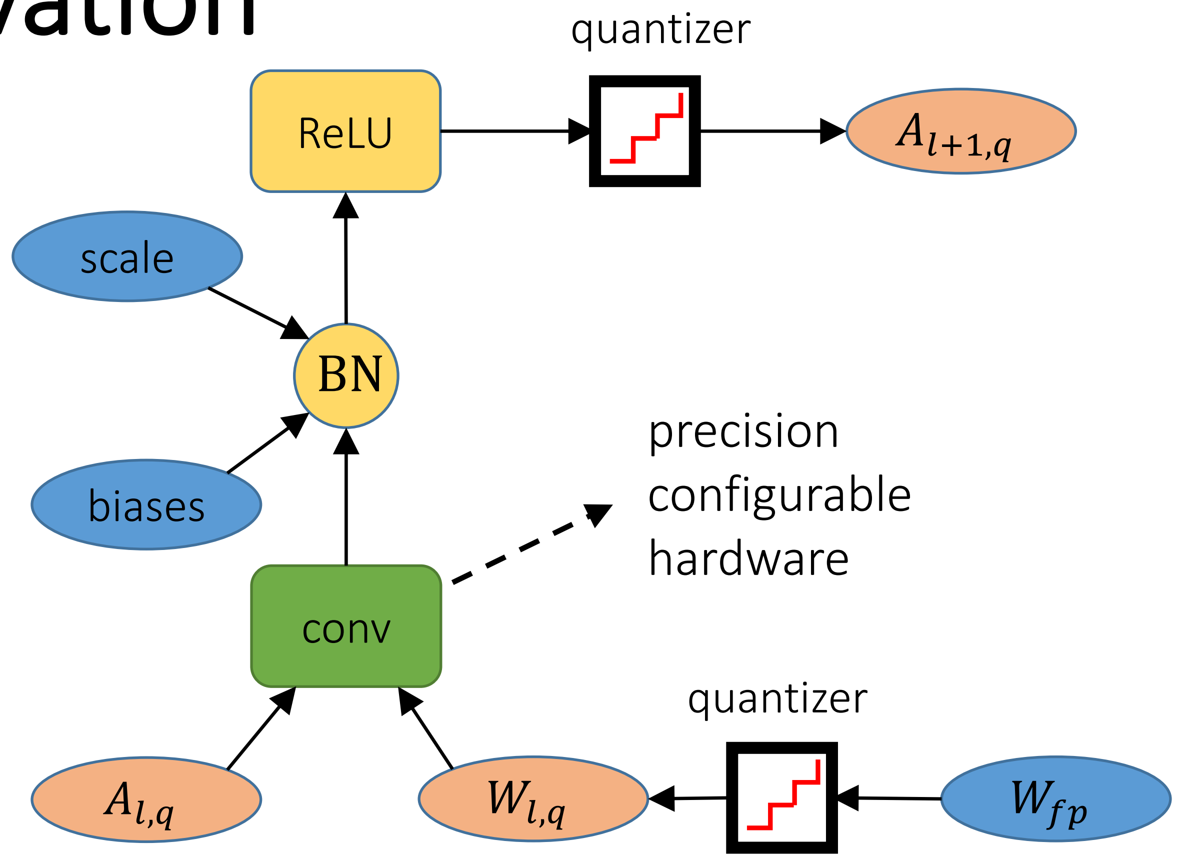


- precision/SNR requirements → changes across layers, datasets and networks
- compact networks more sensitive to quantization

Can we reduce these requirements?

S.K. Gonugondla and N.R. Shanbhag

Motivation



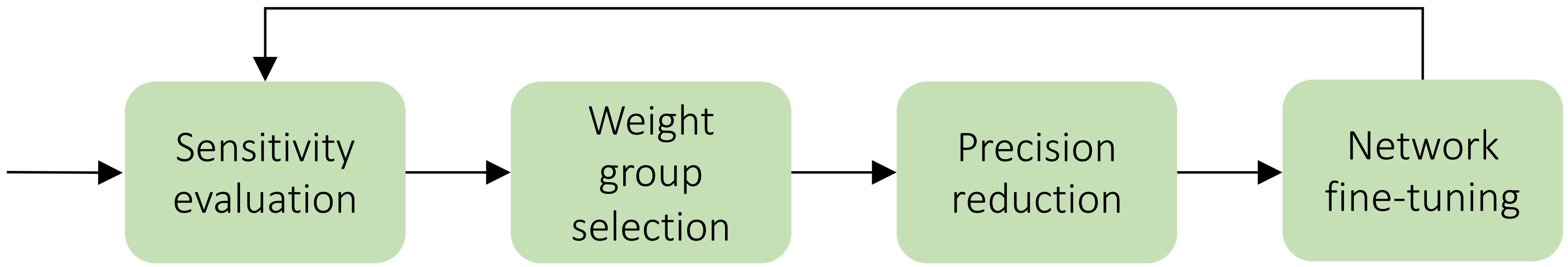
Key insights

- specialized training techniques → aggressively low precision
- granular precision assignments → energy-accuracy trade-off opportunity

Challenge

- granular precision assignments → exponentially huge search space

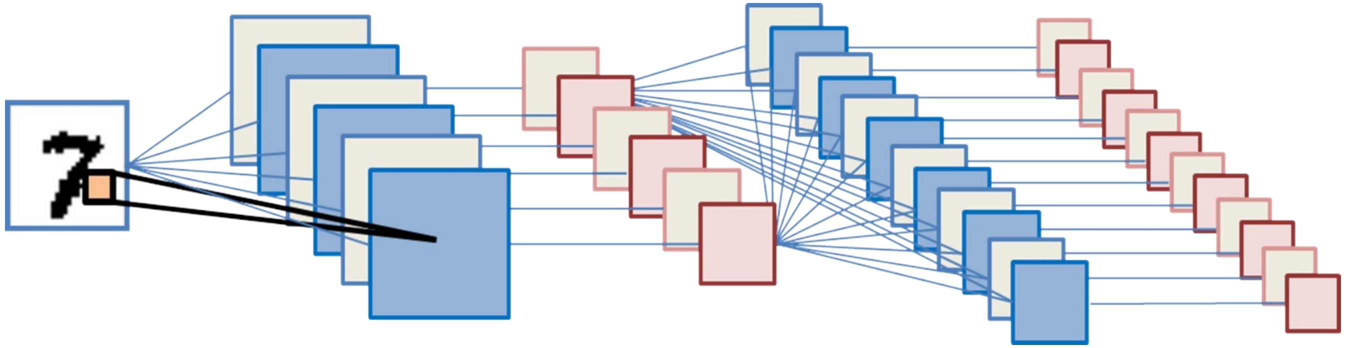
Proposed Scheme

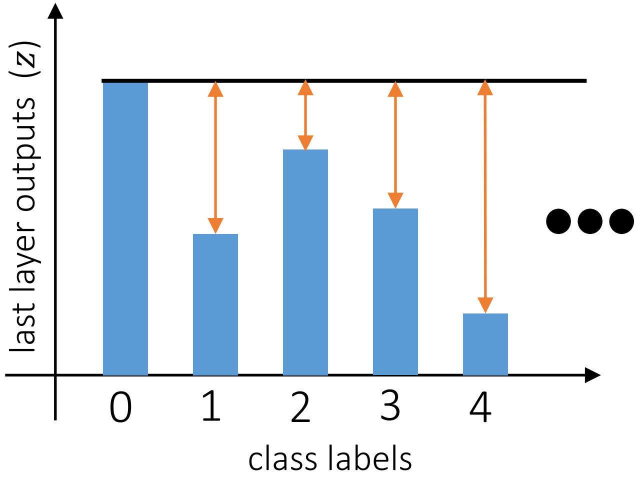


- starting with a pre-trained network ensures a good starting reference
- weight/activation groups that will have the same precision (layer-wise, kernel-wise)
- sensitivity-based precision allocation and retraining
  - protects important weights & compensates for accuracy loss

How do we obtain a sensitivity metric?

Sensitivity Metric



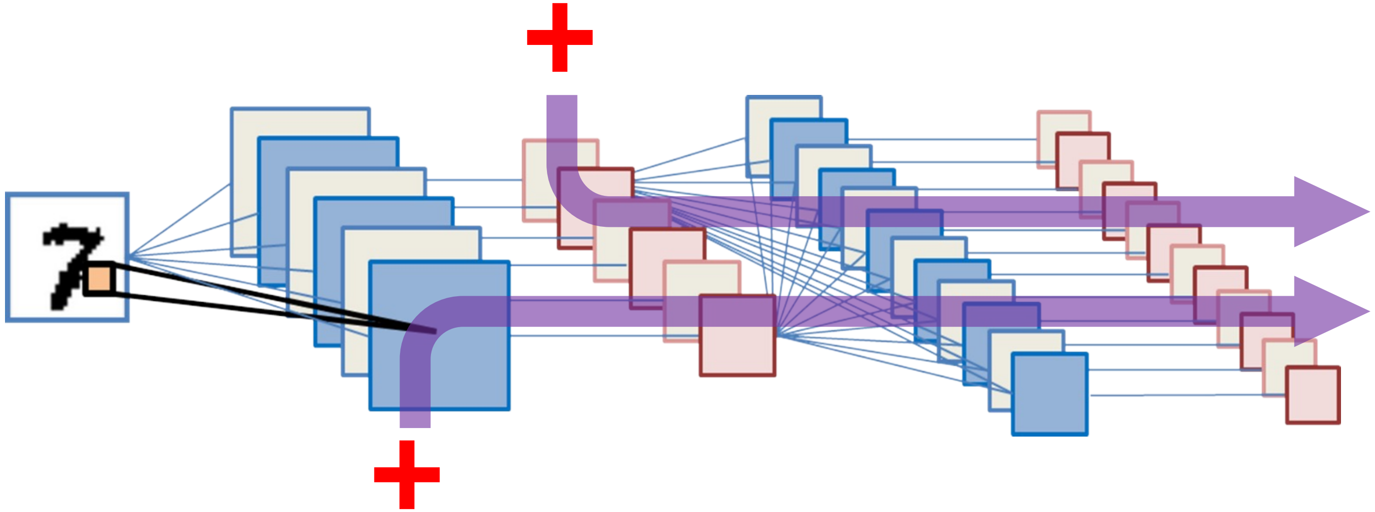


[Sakr *et al.*, ICML-17]

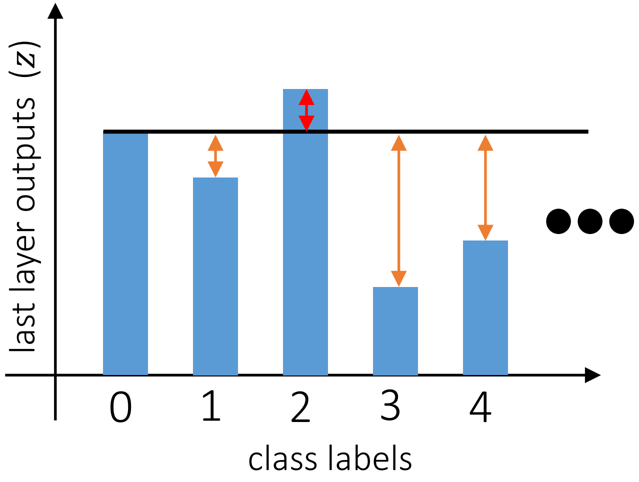
Sensitivity Metric

$$p_m \leq \Sigma_l \Delta_{w,l}^2 E_{w,l} + \Sigma_k \Delta_{a,k}^2 E_{a,k}$$

$\eta \sim \Delta_{B_{w,l}}$ w.p. 0.5



$\eta \sim \Delta_{B_{A,k}}$ w.p. 0.5



$$\Delta_{B_{w,l}} = 2^{-B_{w,l}}$$

$$\Delta_{B_{A,k}} = 2^{-B_{A,k}}$$

l/k : weight/activation group index

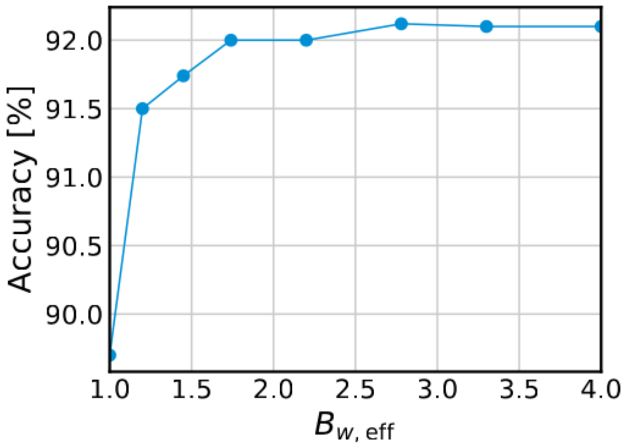
$B_{w,l}$: weight precision of the l-th group

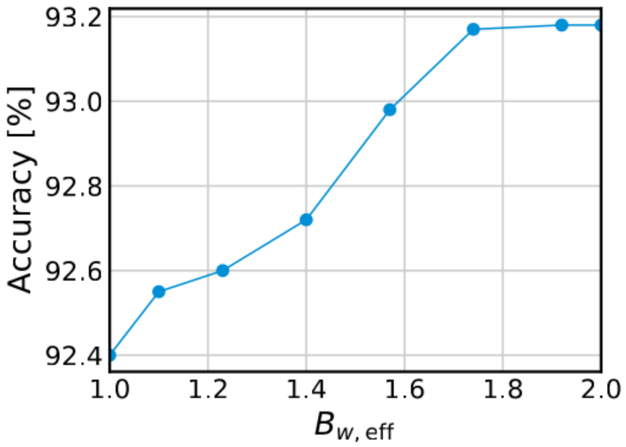
$B_{A,k}$: activation precision of the k-th group

$$E_{W,l} = \mathbb{E}\left[\sum_{\substack{i=1 \\ i \neq \hat{y}_t}}^{M} \frac{\sum_{h \in \mathcal{W}_l} \left|\frac{\partial (Z_i - Z_{y_c})}{\partial h}\right|^2}{12\left|Z_i - Z_{y_c}\right|^2}\right] \qquad E_{A,k} = \mathbb{E}\left[\sum_{i \neq y_c} \frac{\sum_{h \in A_k} \left|\frac{\partial (z_i - z_{y_c})}{\partial h}\right|^2}{12\left|z_i - z_{y_c}\right|^2}\right]$$

[Sakr *et al.*, ICML-17]

Experiments on CIFAR-10



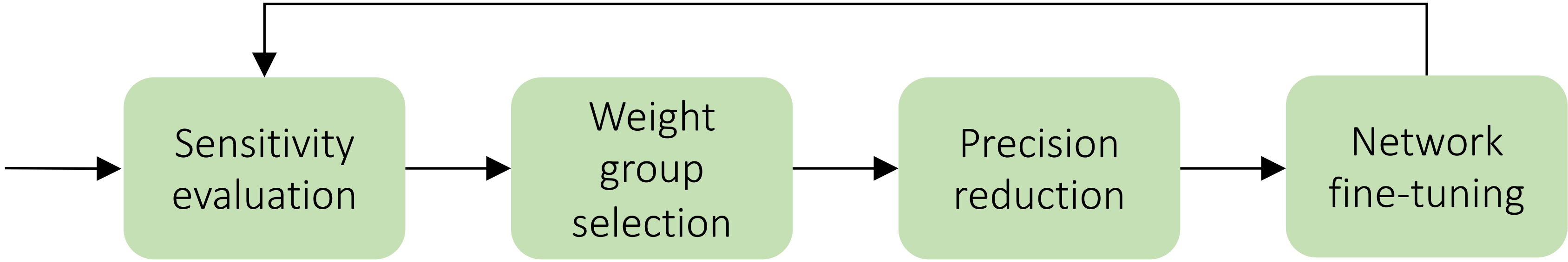


- VGG is less sensitive to quantization than ResNet
- IMPQ achieves high compression with minimal accuracy loss

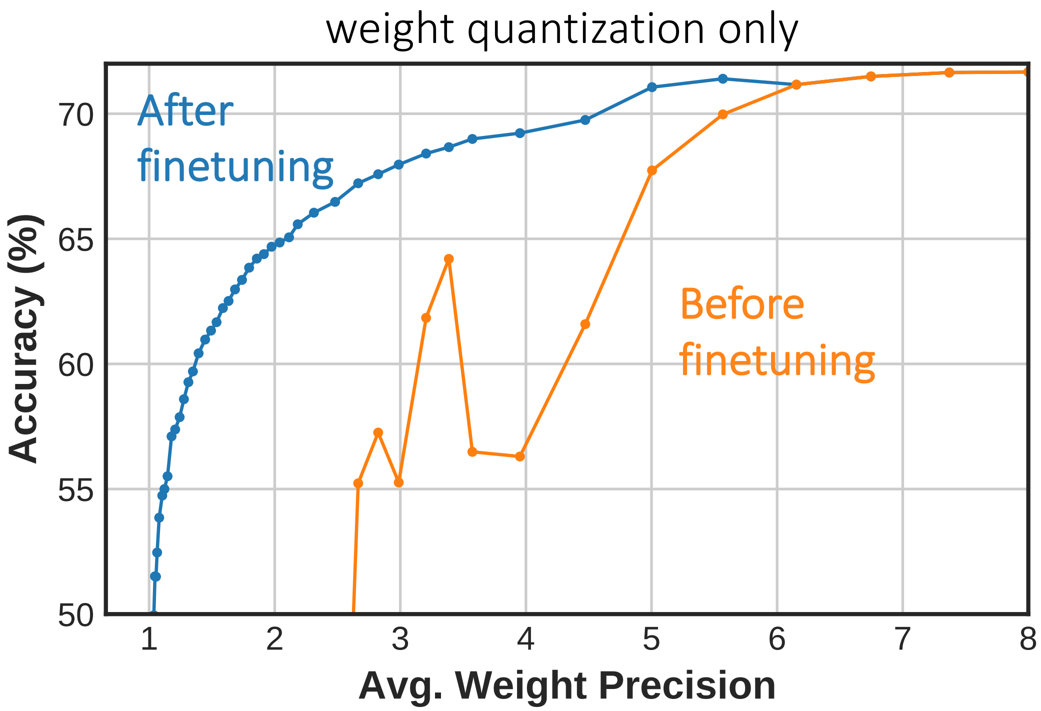
Dataset : CIFAR 10		Network : ResNet-20		
Method	$B_{w,\mathrm{eff}}$	FP† Acc.	Acc. [%]	Change
BWN [26]	1	92.10	90.2	1.90
TWN [6]	Ternary	91.77	90.78	0.89
TTQ [7]	Ternary	91.77	91.13	0.64
ELQ [27]	Ternary	91.25	91.45	-0.20
ELQ [27]	1	91.25	91.15	0.10
DoReFa [9]	3	92.10	91.81	0.29
DoReFa [9]	2	92.10	91.41	0.69
LQ-Net* [25]	3	92.00	92.00	0
LQ-Net* [25]	2	92.00	91.80	0.20
IMPQ	1.74	92.10	92.00	0.10

Dataset : CIFAR 10		Network : VGG-Small		
Method	$B_{w,\mathrm{eff}}$	FP† Acc.	Acc. [%]	Change
BWN [26]	1	93.18	91.77	1.45
TWN [6]	Ternary	93.18	92.56	0.62
LQ-Net* [25]	2	93.8	93.8	0
IMPQ	1.55	93.1	92.97	0.13

Experiments on ImageNet



- ImageNet on MobileNetV1
- 3 epochs of fine-tuning
- kernel-wise granularity
- number of weight groups picked reduced every 8 iterations



S.K. Gonugondla and N.R. Shanbhag
ECE ILLINOIS

Insights : Impact of Kernel-Wise Granularity

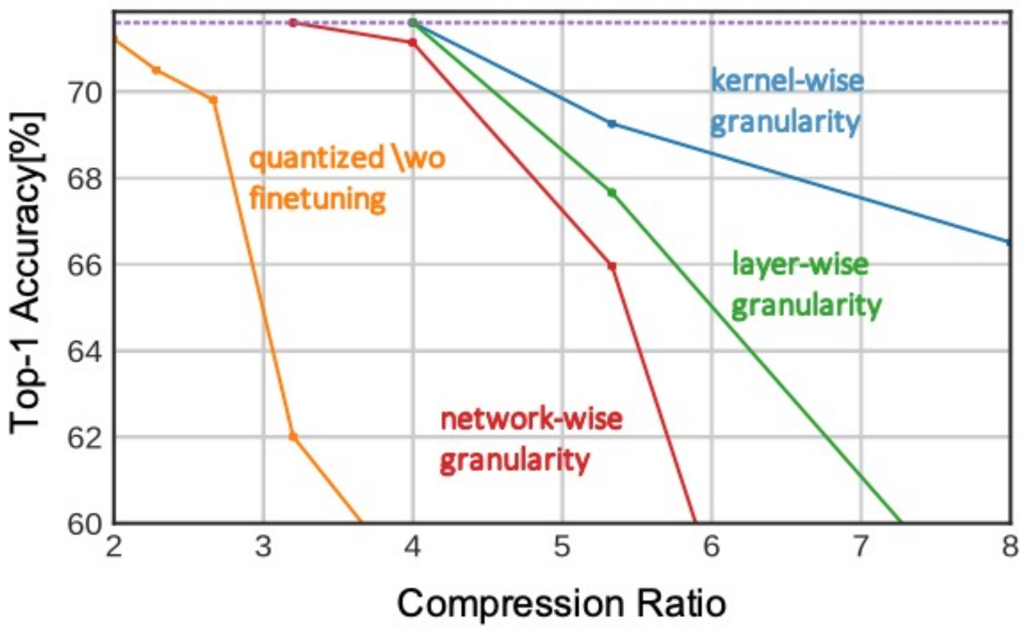
Compression Ratio:

$$CR = \frac{16 \sum N_{w_l}}{\sum N_{w_l} B_{w,l}}$$

w.r.t networks quantized to 16b

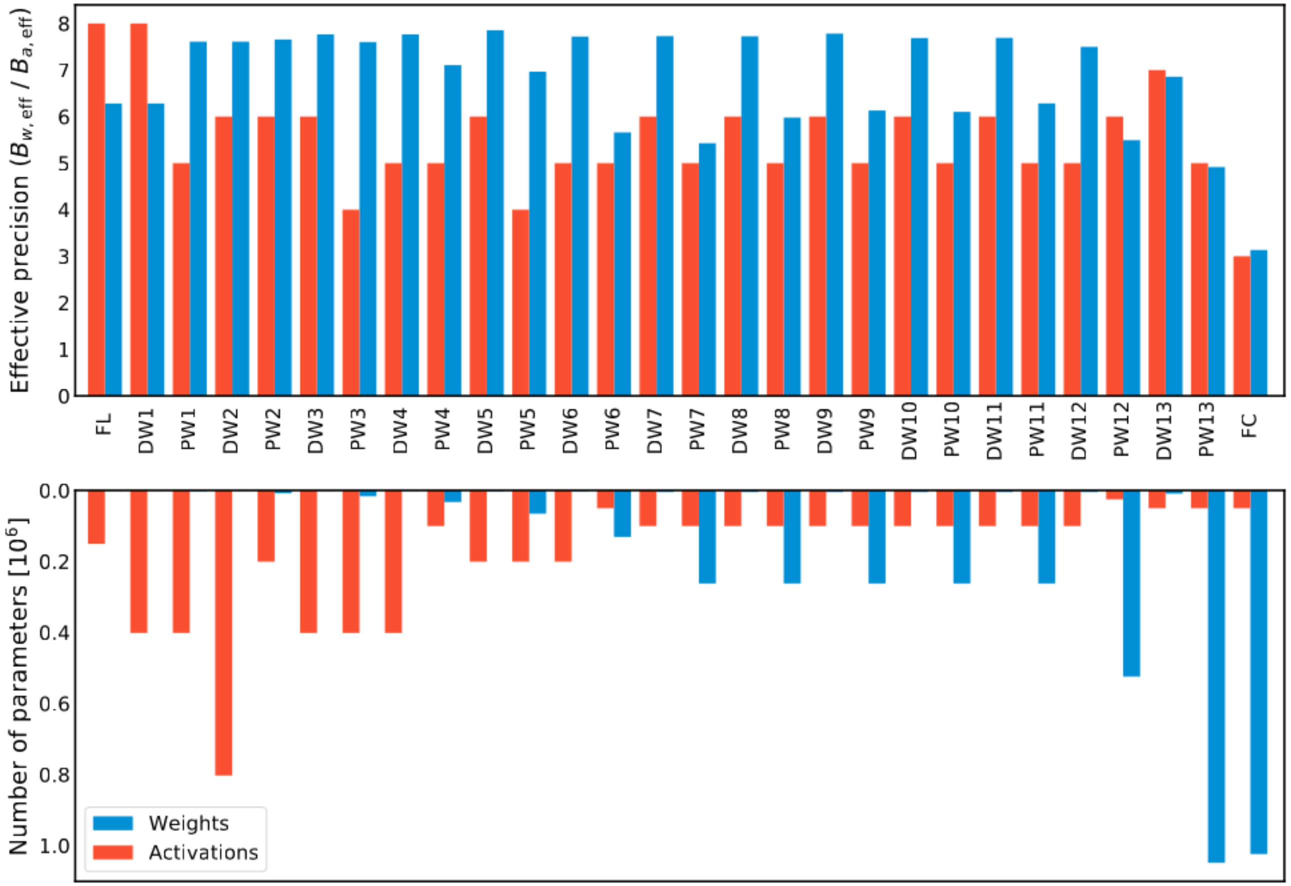
(weight-only quantization)

ImageNet on MobileNetV1



- quantizing a pre-trained network does not lead to large compression
- compression improves with granular precision assignment

Insights: Sensitivity Across Layers

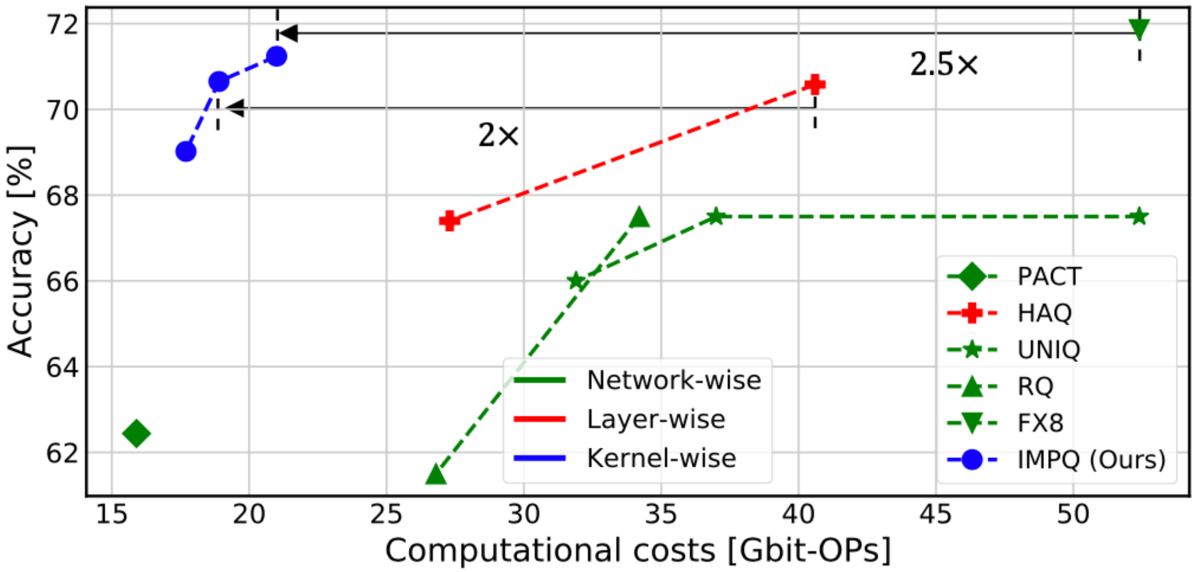


ImageNet on MobileNetV1

- precision requirements reduce with depth
- layers with more parameters are less sensitive to noise
- precision reduced in layers with most parameters

S.K. Gonugondla and N.R. Shanbhag

Comparison With Other Works



Method	$B_{w,\text{eff}}$	$B_{a,\text{eff}}$	Top-1 Acc. [%]	$\mathcal{C}_C$ [Gbit-OPs]
PACT [10, 17]	6	6	71.22	34.2
PACT [10, 17]	5	5	67.00	26.8
PACT [10, 17]	4	4	62.44	15.9
HAQ [17]	6	6	70.90	-
HAQ [17]	5	5	70.58	-
HAQ [17]	4	4	67.40	-
UNIQ [18]	8	8	67.50	52.4
UNIQ [18]	5	8	67.50	37.0
UNIQ [18]	4	8	66.00	31.9
RQ [19]	6	6	67.50	34.2
RQ [19]	5	5	61.50	26.8
DBQ* [8]	3	8	70.92	21.8
FP Baseline	32	32	71.84	-
FX8 Baseline	8	8	71.86	52.4
IMPQ	6	6	71.24	21.0
IMPQ	5	5	70.65	18.9
IMPQ	4	5.8	69.02	17.7

* nonlinear quantization

- IMPQ reduces costs by 2×—2.5× on MobileNetV1
- 1.7× better compression

Conclusions

- Granular precision assignment **leads to lower precision but** is challenging to implement
- Proposed method uses sensitivity-based precision reduction
- 42% better compression compared to s.o.t.a on CIFAR-10
- 33% better compression on MobileNet-V1
- <1.5% drop in accuracy for $B_{a,\mathrm{eff}} = B_{w,\mathrm{eff}} = 5$ on MobileNet-V1

Thank You!

gonugon2@Illinois.edu
shanbhag@Illinois.edu

Acknowledgement: This work was supported by the Defense Advanced Research Agency (DARPA) and the Semiconductor Research Corporation (SRC) via the Center for Brain-inspired Computing (C-BRIC) and the AI Hardware (AIHW) program.