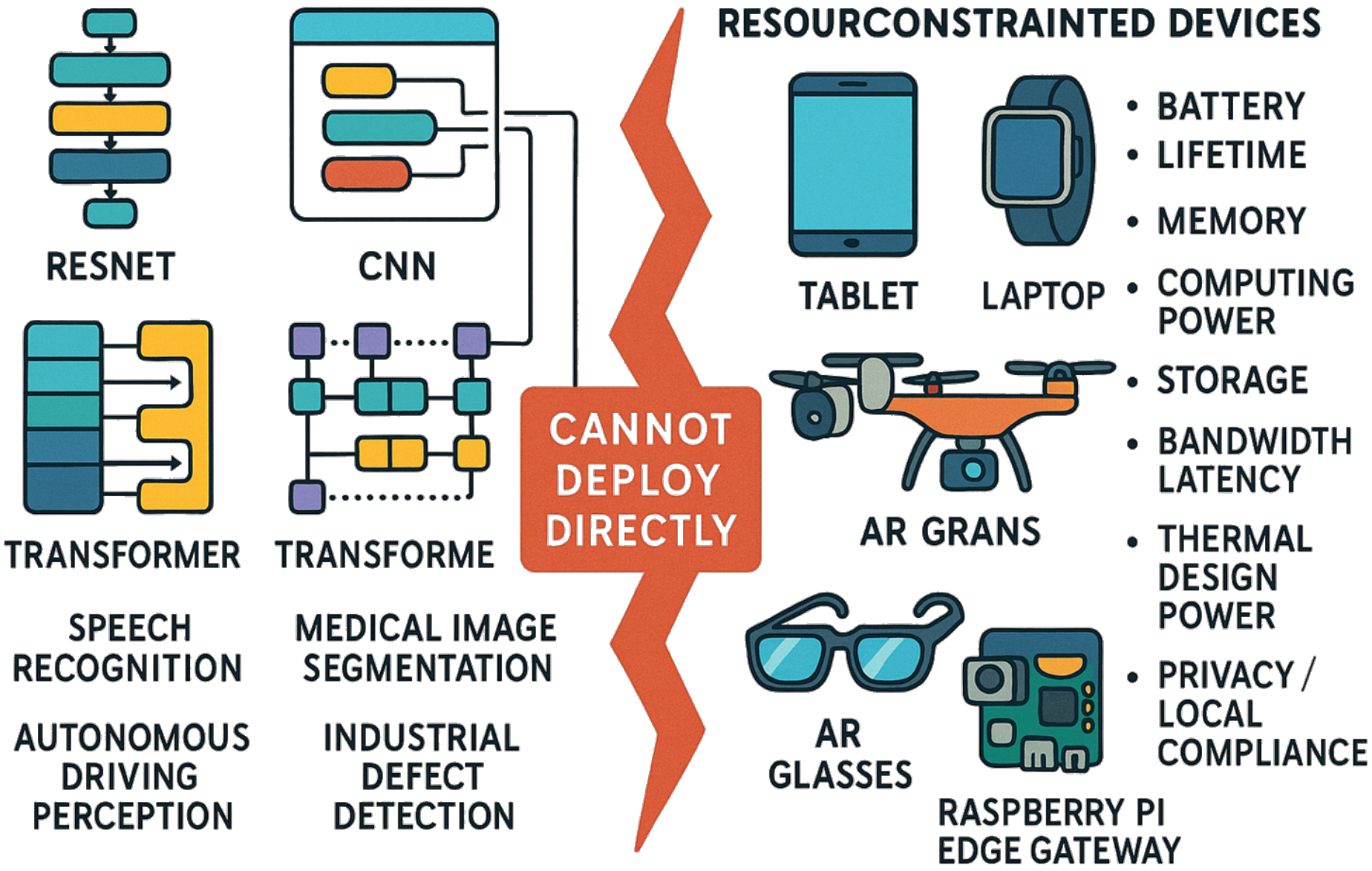
Efficient and Generalizable Mixed-Precision Quantization via Topological Entropy NeurIPS-2025

Authors: Nan Li, Yonghui Su, Lianbo Ma

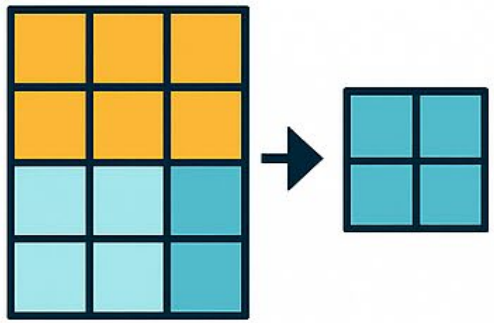
E-mails: linan10@sxu.edu.cn; 2371400@stu.neu.edu.cn; malb@swc.neu.edu.cn

Deep Models and Resource-Constrained Devices



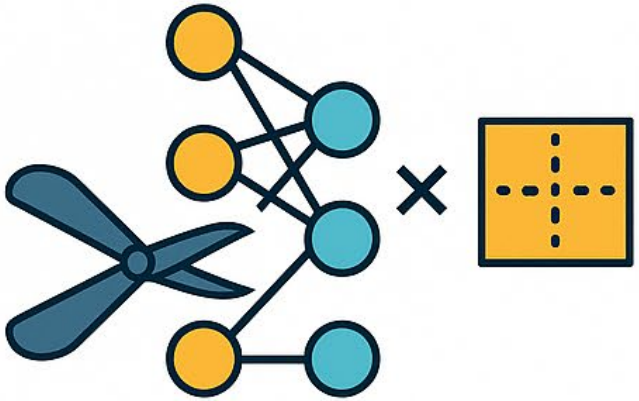
Model Compression

QUANTIZATION



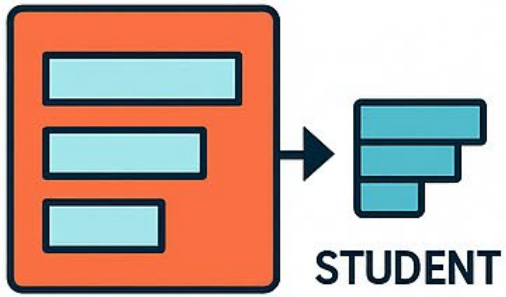
Reduces the precision of model weights and activations

PRUNING



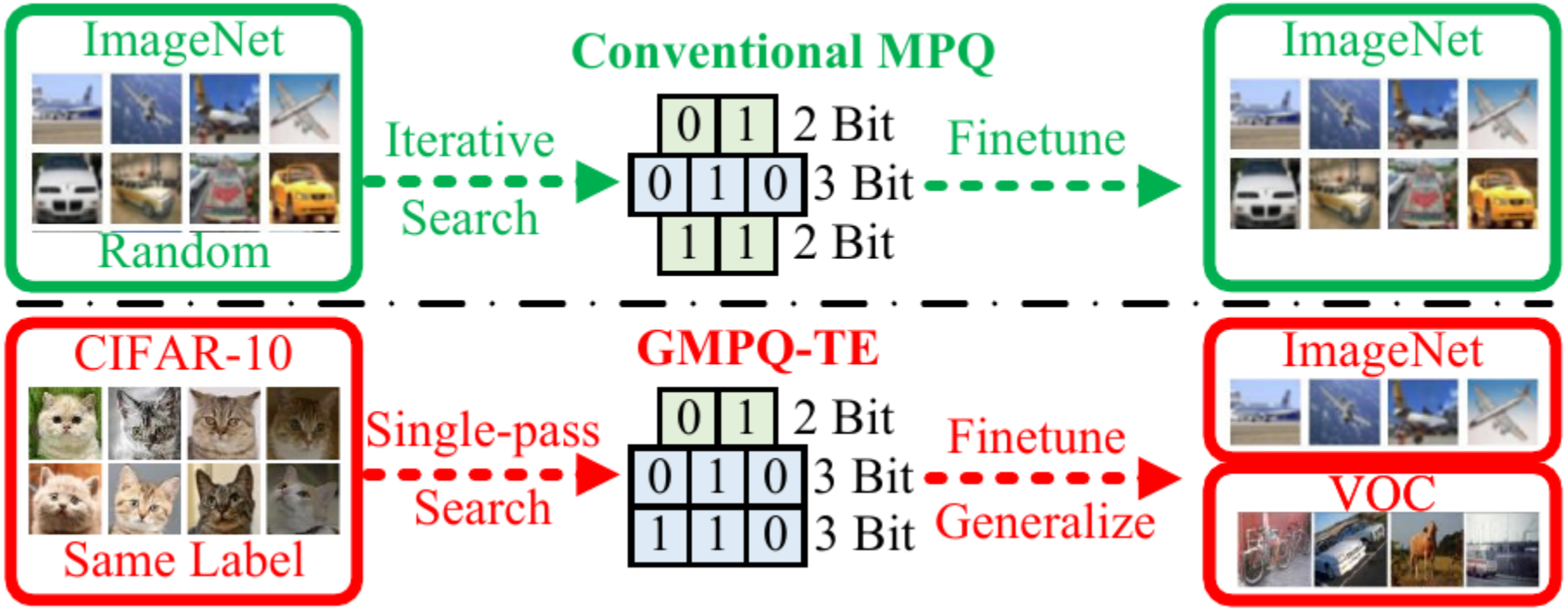
Removes unnecessary model weights and connections

KNOWLEDGE DISTILLATION



Trains a smaller model to mimic a larger model

Mixed-Precision Quantization



Comparison of the search cost used to obtain the optimal quantization policy on MobileNet-V2 between GMPQ-TE and other MPQ methods.

Contributions

- We first attempt to use TE for measuring the quantization sensitivity of a network model or a layer by a minibatch of data with the same label.

- We provide a comprehensive theoretical analysis of the performance degradation boundary and resolution– and label–independent of TE.

- We conduct extensive experimental results on image classification and object detection, which show that the quantization policy generalized from CIFAR-10 to ImageNet and PASCAL VOC achieves a competitive accuracy-complexity trade-off.

Methodology

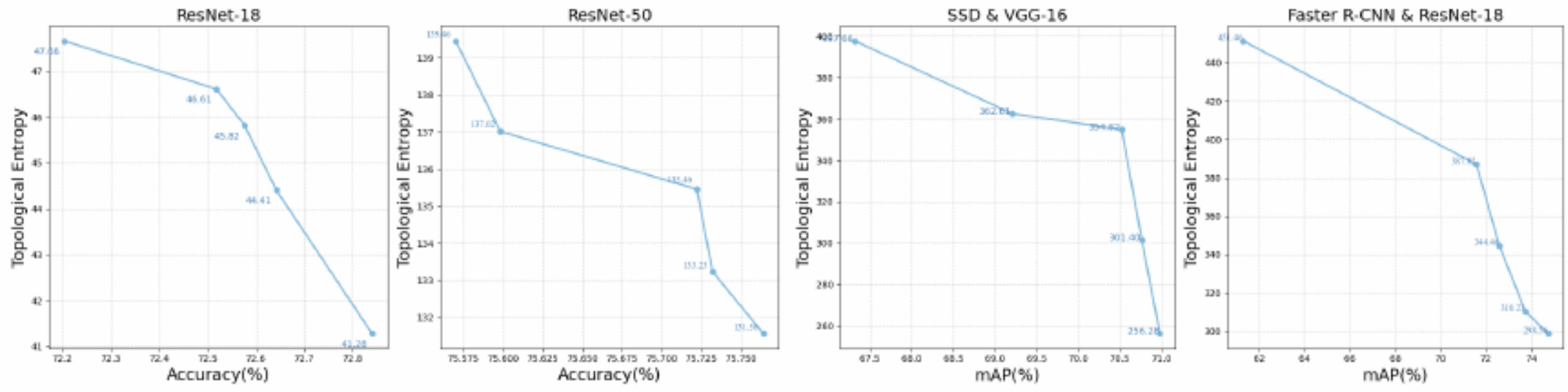
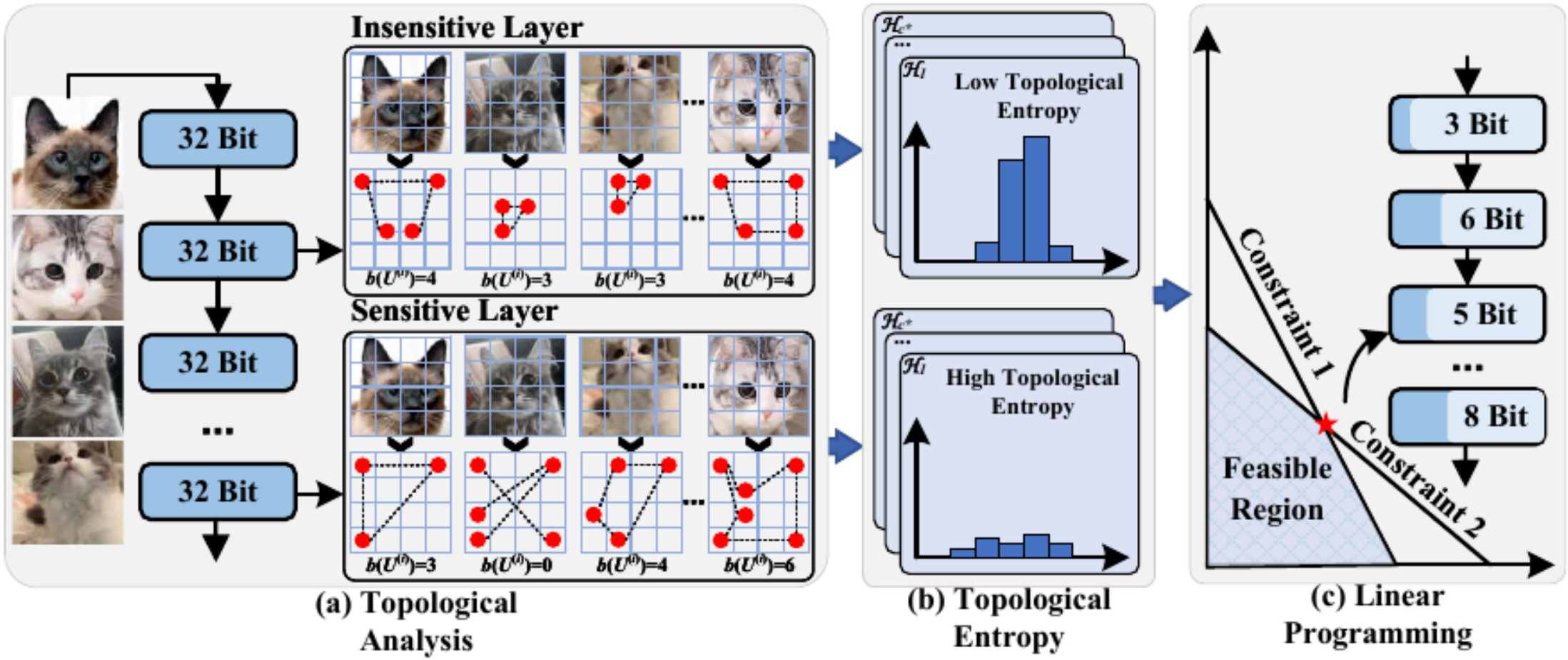


Figure 4: Relationship between TE and accuracy/mAP for different quantization policies on ResNet-18, ResNet-50, SDD & VGG-16 and Faster R-CNN & ResNet-18.

Methodology



The overall framework of the proposed GMPQ-TE method. (a) Analyze a minibatch of data with the same label and obtain Betti time. (b) Calculate TE based on Betti time distribution. (c) Linear programming problem constructed by TE to derive optimal generalizable quantization policy.

Experiments and Results

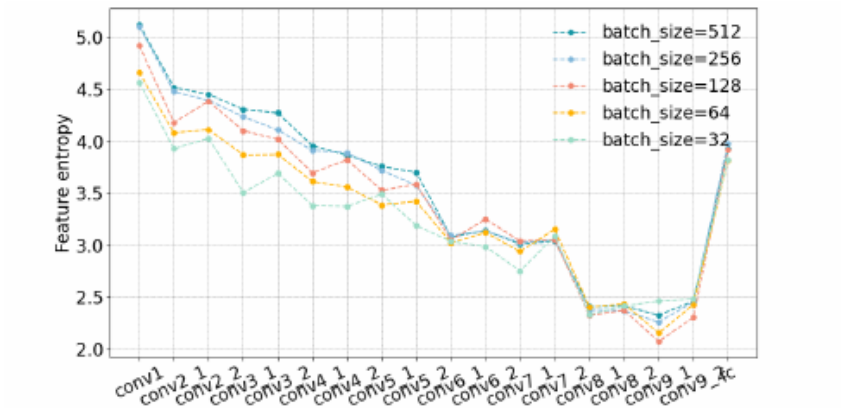


Figure 5: Relationship between different batch sizes and TE on ResNet-18.

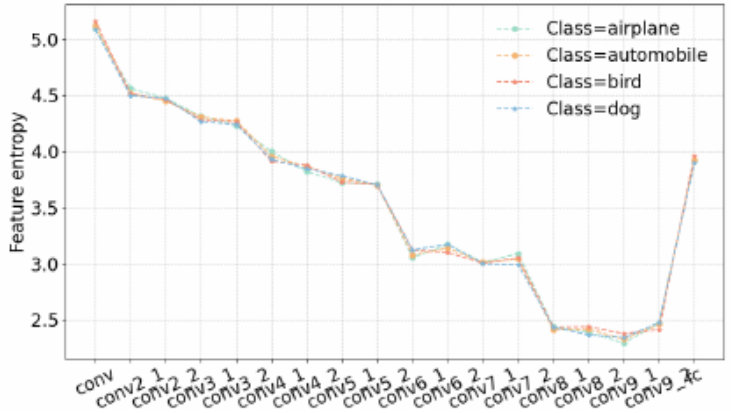
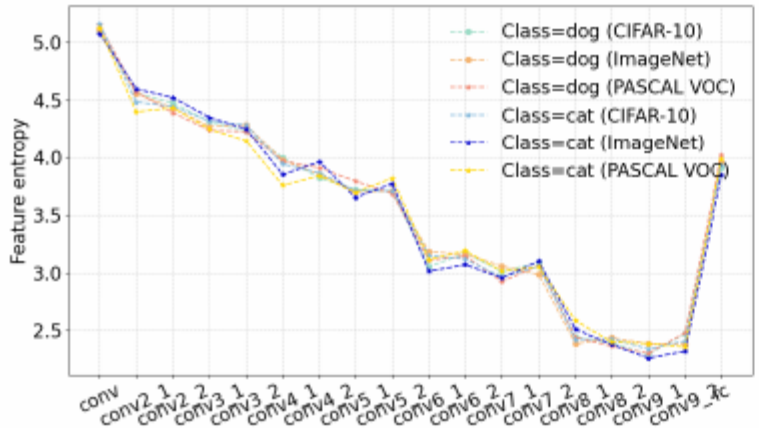
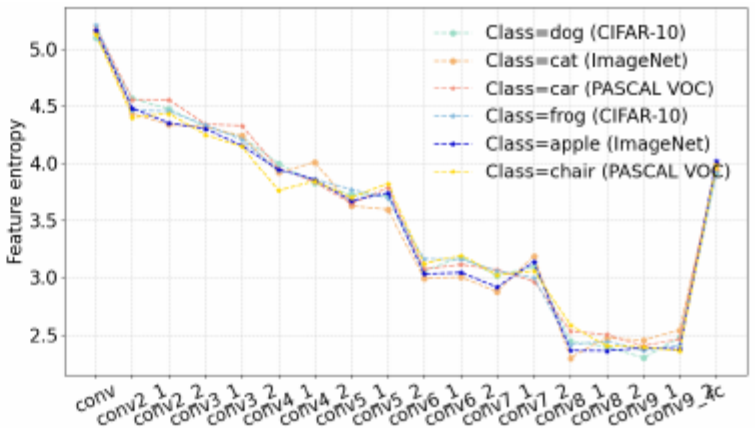


Figure 6: Relationship between different labels from the same dataset and TE on ResNet-18.



(a) Same label images



(b) Different label images

Figure 7: Relationship between labels from different datasets and TE on ResNet-18.

Experiments and Results

Table 1: Results for image classification on MobileNet-V2 (PTQ).

Methods	Params.	BOPs	Comp.	Top-1	Top-5	Cost.
Full-precision	13.4	337.9	–	71.9	90.3	–
RQ	2.7	11.9	28.4	68	–	–
GMPQ†	1.4	10.4	32.6	71.5	90.1	6.1K
R-GMPQ†	2.2	9.9	34.1	71.7	90.2	6.8K
GMPQ-TE	1.4	10.1	33.4	71.8	90.2	15
HAQ	1.4	8.3	41	69.5	88.8	18.3K
DJPQ	1.9	7.9	43	69.3	–	43.9K
GMPQ	1.2	7.4	45.8	70.4	89.4	9.3K
R-GMPQ	1.1	7.2	46.7	70.9	89.7	10.4K
GMPQ-TE	1.1	7.3	46.1	71.2	89.9	14
HMQ	1.7	5.2	64.4	70.9	–	120.6K
DQ	1.7	4.9	68.7	69.7	–	77.7K
EMQ	1.5			70.75	–	few seconds
GMPQ†	1	4.8	69.7	69.5	89.1	10K
R-GMPQ†	1.3	4.7	72.2	69.7	89.5	11.1K
OMPQ*	1.5	12.3	27.3	70.39	89.9	9
GMPQ-TE	1.3	4.5	74.8	69.8	89.7	11

Table 2: Results for object detection on SSD & VGG-16. (PTQ)

Methods	Params.	BOPs	Comp.	mAP	Cost.
Full-precision	105.5	27787.7	–	72.4	–
HAQ	42.7	847.2	32.8	70.9	225K
HAQ-C	42.9	819.7	33.9	67.6	18.3K
EdMIPS	33.5	958.2	29	69.4	5.4K
GMPQ	36.6	796.2	34.9	70.5	5.7K
R-GMPQ	32.6	761.3	36.5	70.8	6.4K
GMPQ-TE	32.6	758.4	36.6	71.1	27
HAQ	35.5	430.15	64.6	69.1	244.4K
HAQ-C	36.3	445.3	62.4	66.4	24.4K
EdMIPS	29.4	454	61.2	68.7	108.7K
GMPQ†	24.7	413.5	67.2	69.2	6.4K
R-GMPQ†	26.9	406.8	68.3	70.3	7.2K
GMPQ-TE	24.7	405.4	68.6	70.6	23

Table 3: Comparison of different quantization methods on two FPGA boards for ResNet-18 on ImageNet.

Methods	Results on FPGA XC7Z020				Results on FPGA XC7Z045			
	Utilization		Throughput (GOP/s)	Latency (ms)	Utilization		Throughput (GOP/s)	Latency (ms)
	LUT	DSP			LUT	DSP		
GMPQ	49%	100%	75	57.3	52%	100%	367	12.3
EdMIPS	39%	100%	61	61.4	41%	100%	316	15.9
R-GMPQ	50%	100%	79	52.1	59%	100%	398	10.4
GMPQ-TE	53%	100%	84	42.3	62%	100%	413	9.4

Thanks for listening.

Authors: Nan Li, Yonghui Su, Lianbo Ma

E-mails: linan10@sxu.edu.cn; 2371400@stu.neu.edu.cn; malb@swc.neu.edu.cn