Entropy-Driven Mixed-Precision Quantization for Deep Network Design

Zhenhong Sun, Ce Ge, Junyan Wang, Ming Lin,
Hensen Chen, Hao Li and Xiuyu Sun

Zhenhong Sun
Algorithm Engineer
DAMO Academy, Alibaba Group
E:zhenhongsun1992@outlook.com

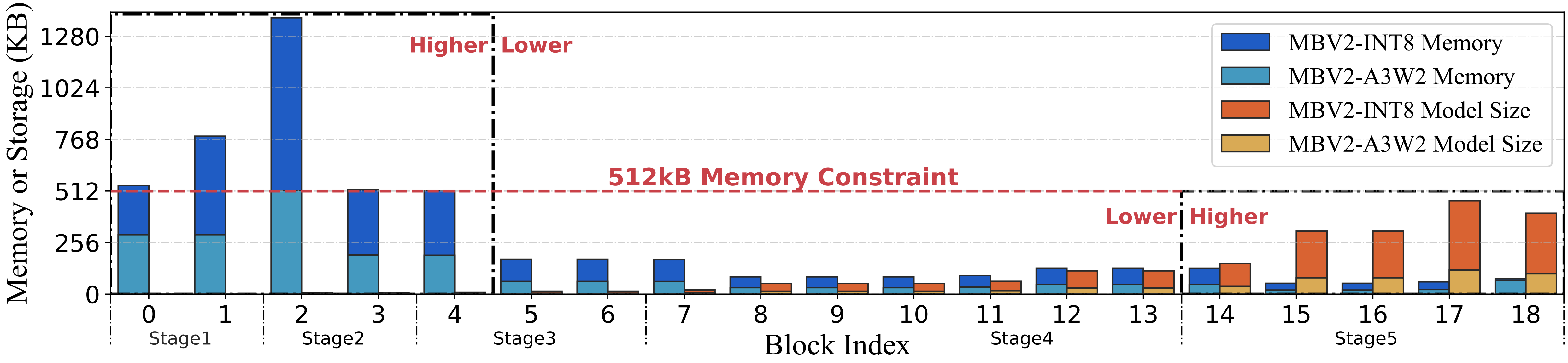
Contact Xiuyu by DingTalk

Contact Xiuyu by WeChat

Outline

- ✓Motivation
- ✓Quantization Entropy Score
  - Overall
  - Full-precision model
  - Mixed-precision model
  - Gaussian Calibration
  - QBR
- ✓Experimental Results
- ✓Conclusion

Motivation



✓On IoT devices, **limited SRAM memory and Flash storage** are the major constraints for deploying models. A SOTA ARM Cortex-M7 MCU merely has **512kB SRAM and 2MB Flash**.

✓Typically, full-precision MobileNetV2 consumes 5.6M peak memory and 13.5M storage, and INT8 model consumes 1.4M peak memory and 3.4M storage.

✓The resource utilization of fixed-precision quantization is not high in some stages.

Motivation

Table 1: TOP-1 ACC of fixed-precision MobileNetV2 models on ImageNet with 120 training epochs. Bold values meet the 512KB SRAM limit, and underline values meet the 2MB Flash limit.

Activation Bit	Weight Bit 2	3	4	5	6	8
3	**47.43**	**59.38**	**62.78**	**63.59**	**64.06**	**64.28**
4	55.54	64.74	67.78	68.50	68.59	68.94
5	56.66	66.31	69.23	69.75	69.99	70.25
6	57.73	66.62	69.11	70.00	70.07	70.48
8	57.89	66.69	69.25	70.02	70.17	70.60

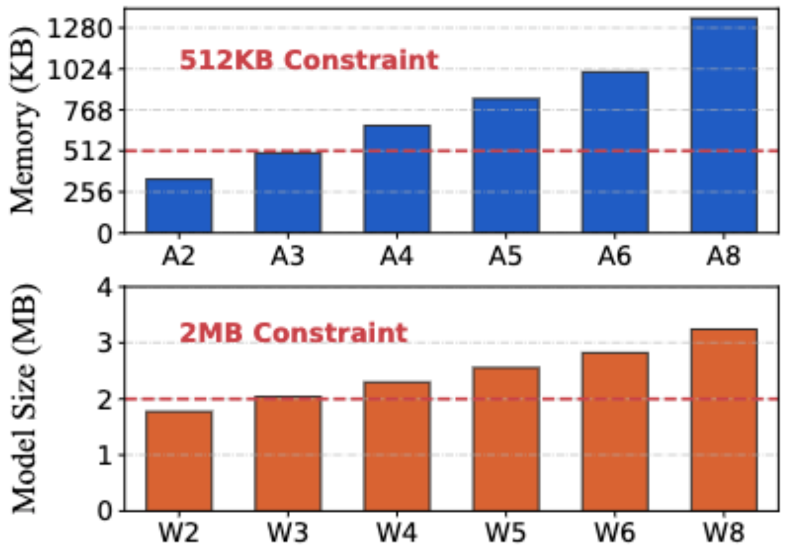
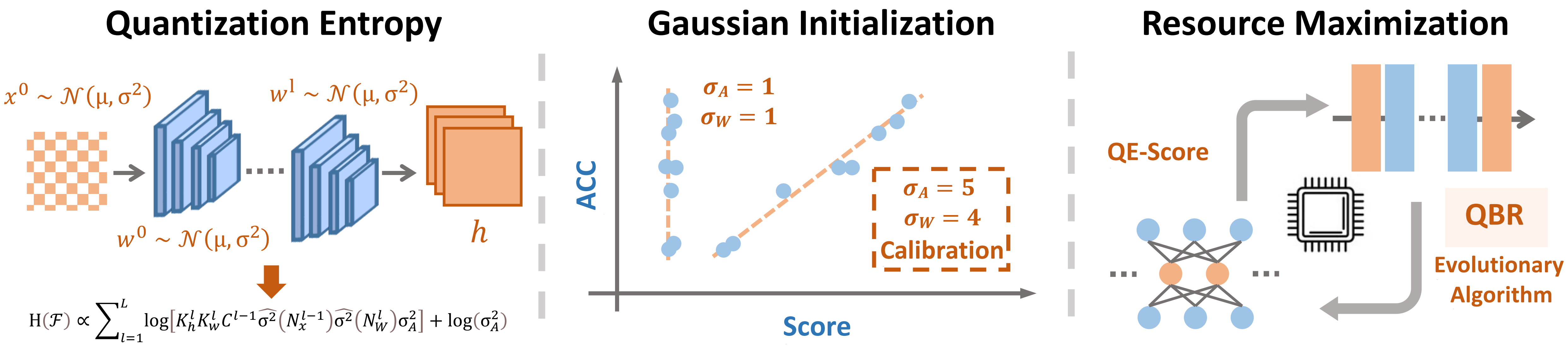


Figure 3: Peak memory and model size of fixed-precision MobileNetV2 models.

- ✓ Thus, deploying model on IoT devices needs much lower bit precision quantization.
- ✓ **Lower bit precision lower accuracy**. “A3W2” model fits both 512KB memory limit and 2MB storage limit.
- ✓ Mixed-precision quantization with NAS can use **lower bit on tight-resource position and higher bit on rich-resource position** to promote the deloyment.

Quantization Entropy--Overall



✓Our proposed strategy for deep network design consists of three modules, including quantization entropy score, Gaussian initialization calibration, and resource maximization.

Quantization Entropy--Full-Precision Models

✓Regard a deep neural network as an information system, and the differential entropy of the last output feature map represents the expressiveness.

✓The differential entropy of a Gaussian distribution only depends on Variance.

$$H(x) = \int_{-\infty}^{+\infty} -\log(p(x))p(x)\,dx \quad \propto \log(\sigma^2), \tag{1}$$

✓The L-layer's expectation and variance without quantization:

$$\mathbb{E}(\boldsymbol{x}_i^l) = 0, \quad \sigma^2(\boldsymbol{x}_i^l) = \sum_{h=1}^{K_h^l} \sum_{w=1}^{K_w^l} \sum_{c=1}^{C^{l-1}} \left[\sigma^2(\boldsymbol{x}_{chw}^{l-1}) \times \sigma^2(\boldsymbol{W}_{chw}^l)\right]. \tag{4}$$

✓Next, we will explore how to introduce mixed-precision quantization into the calculation process.

Quantization Entropy--Mixed-Precision Models

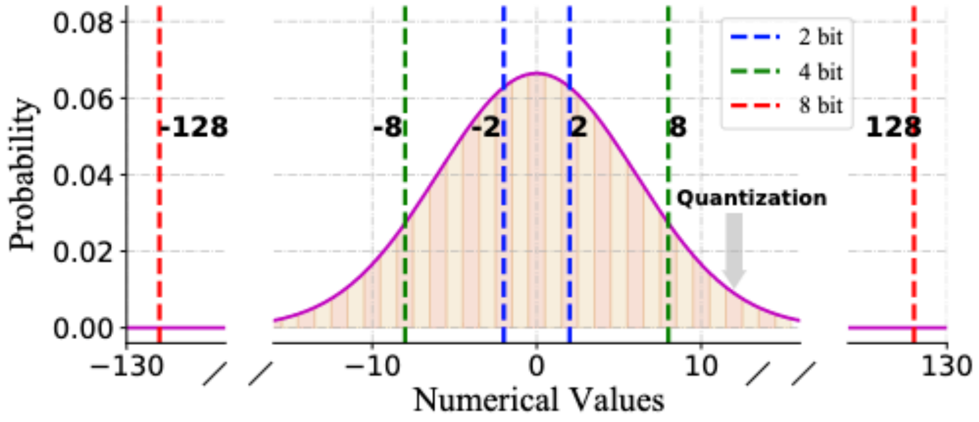


Figure 4: $N = \{2, 4, 8\}$ bit quantization on Gaussian variable. The upper and lower bounds represent truncation. Shaded areas represent quantization.

Table 2: Look up table of $\hat{\sigma}(N)$ according to σ and N bit. Low-precision leads to small variance.

σ	N Bit Precision 2	3	4	5	6	7	8
1	1.00	1.04	1.04	1.04	1.04	1.04	1.04
2	1.47	1.94	2.02	2.02	2.02	2.02	2.02
4	1.73	2.89	3.85	4.01	4.01	4.01	4.01
6	1.82	3.26	5.04	5.96	6.00	6.00	6.00

- ✓We need to insert low-precision function behind Gaussian initialized input and weights.
- ✓According to Fig. 4, the quantization of Gaussian variable will decrease the variance of the input, which is the reason of quantization loss.
- ✓The decreased quantization standard deviation ˆσ(N) is demonstrated in Table 2.

Quantization Entropy--Mixed-Precision Models

✓Thus we consider to refine entropy score to distinguish the different bits loss. We adopt a scaling parameter to normalize the input to σ_A :

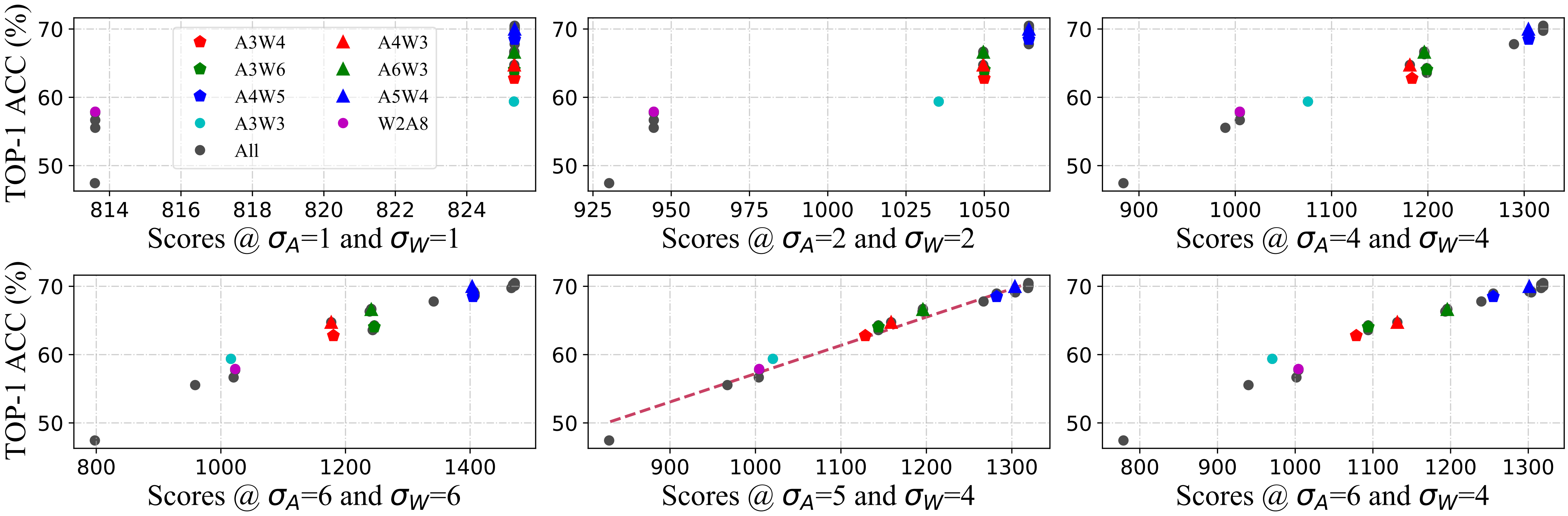
$$\sigma^2(\boldsymbol{x}_i^l) = \sum_{h=1}^{K_h^l} \sum_{w=1}^{K_w^l} \sum_{c=1}^{C^{l-1}} \left[\hat{\sigma}^2(N_x^{l-1}) \times \hat{\sigma}^2(N_W^l)\right] \times \sigma_S^2(\boldsymbol{x}_{chw}^{l-1}), \quad \sigma_S^2(\boldsymbol{x}_{chw}^{l-1}) = \sigma^2(\boldsymbol{x}_{chw}^{l-1})/\sigma_A^2. \quad (7)$$

✓Finally form the Quantization Entropy Score (QE-Score) to measure quantization loss:

$$H(\mathcal{F}) \propto \sum_{l=1}^{L} \log\left[K_h^l K_w^l C^{l-1} \hat{\sigma}^2(N_x^{l-1}) \hat{\sigma}^2(N_W^l)/\sigma_A^2\right] + \log(\sigma_A^2), \qquad (9)$$

✓QE-Score depends on the structural parameters, quantization precision, and initial standard deviation σ_A and σ_W. Next, we will show how to determine σ_A and σ_W.

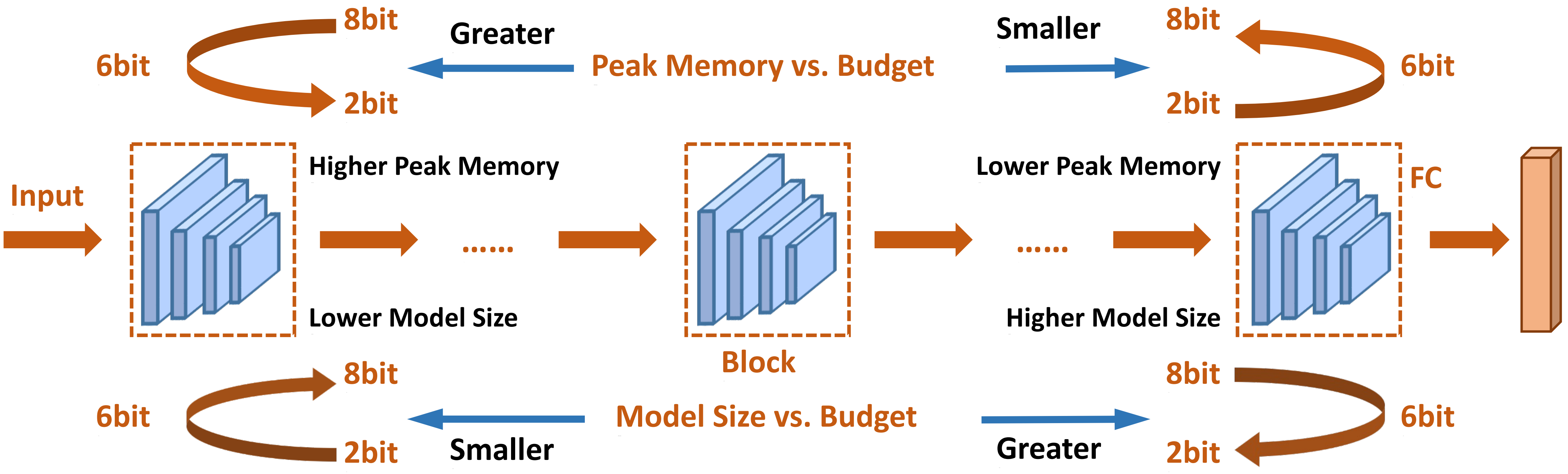
Quantization Entropy-- Gaussian Calibration



✓Use MobileNetV2 architecture with fixed-precisions to calibrate the accuracy and score.

✓When gradually increasing the values of σ_A and σ_W , QE-Scores are gradually positive correlated with accuracy. $\sigma_A = 5$ and $\sigma_W = 4$ can rank the diversity of activations and weights on accuracy.

Quantization Entropy--QBR

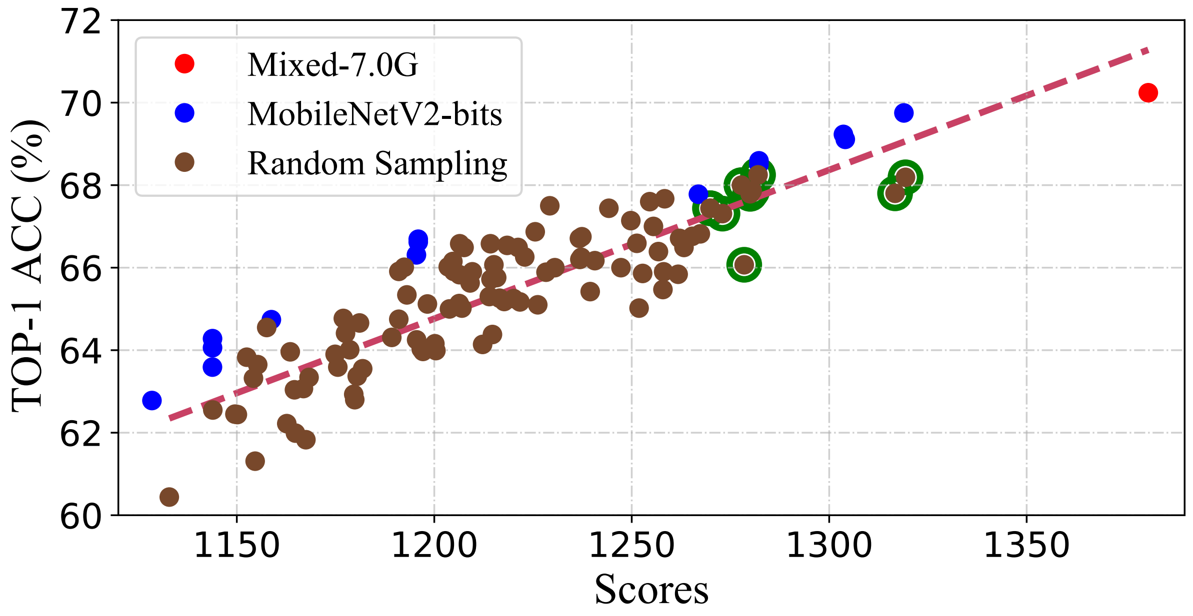


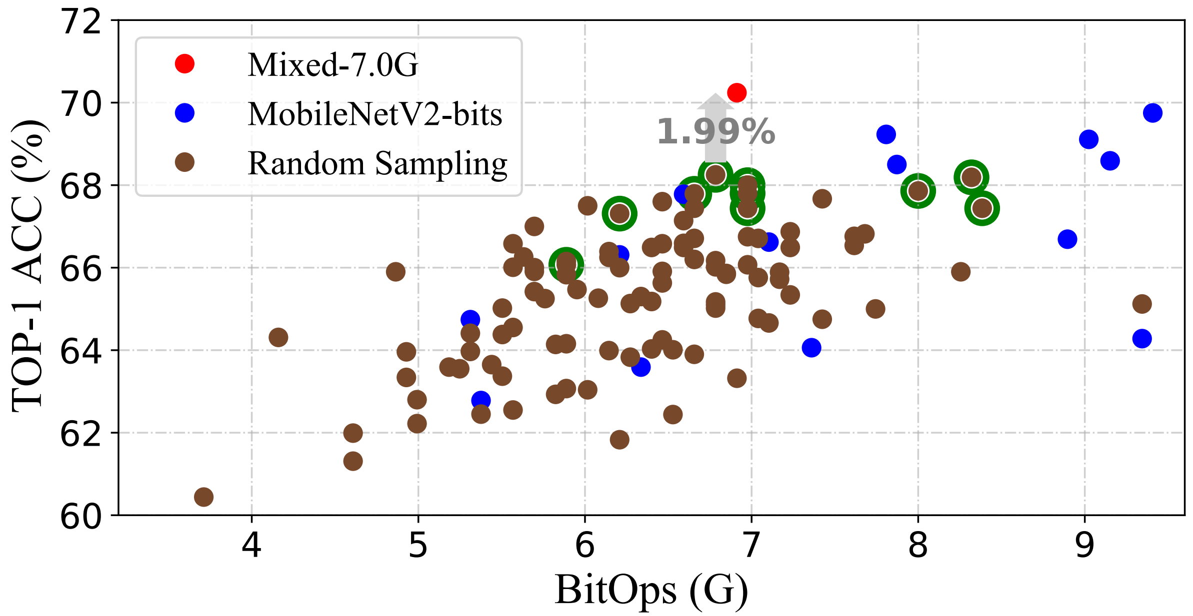
✓Finally, to maximize the resource utilization, we use Quantization Bits Refinement (QBR) strategy to redistributes the mixed-precisions.

✓Given a cadidate structure, we scale the mixed-precision of activations to make the peak memory meet the budget. Accordingly, we also increase or decrease the mixed-precision of weights to guarantee the model size approaches the budget.

Experimental Results – Correlation study

Alibaba Group 阿里巴巴集团





✓To verify the correlation between our QE-Score and accuracy, we randomly selected 100 models without QE-Score under the same searching space.

✓Figures show Our QE-Score is valid to rank various architectures without training.

Experimental Results – Mixed Quantization Comparison

Table 3: Comparison with state-of-the-art efficient models with mixed-precision quantization. MBV2-4bit use 4-bit for the overall layers except for the first and last layer. †: 64 cores of Intel(R) Xeon(R) Platinum 8269CY CPU @ 2.50GHz.

Model	Quant.	Search Devices	Design Cost (hours)	Model Size (MB)	BitOps (G)	ImageNet TOP-1	CO2e (marginal)
MBV2 [28]	8-bit	-	-	3.4	19.2	71.9%	-
MBV2 [28]	4-bit	-	-	2.3	7.0	68.9%	-
MBV2+HAQ [34]	mixed	GPUs	96N	-	-	71.9%	27.23N
DNAS [36]	mixed	GPUs	300N	-	57.3	74.0%	11.34N
SPOS [11]	mixed	GPUs	288+24N	-	51.9	74.6%	82+6.81N
APQ [35]	mixed	GPUs	2400+0.5N	-	16.5	74.1%	672+0.14N
APQ [35]	mixed	GPUs	2400+0.5N	-	23.6	75.1%	672+0.14N
Ours-19.2G	mixed	CPUs†	0.5N	3.2	18.8	**74.8%**	0.19N
Ours-7.0G	mixed	CPUs†	0.5N	2.2	6.9	**70.8%**	0.19N

✓ Our searched model has a higher accuracy boost than MobileNetV2 baseline, which is also better than other methods.

Experimental Results – Classification on ImageNet

Alibaba Group 阿里巴巴集团

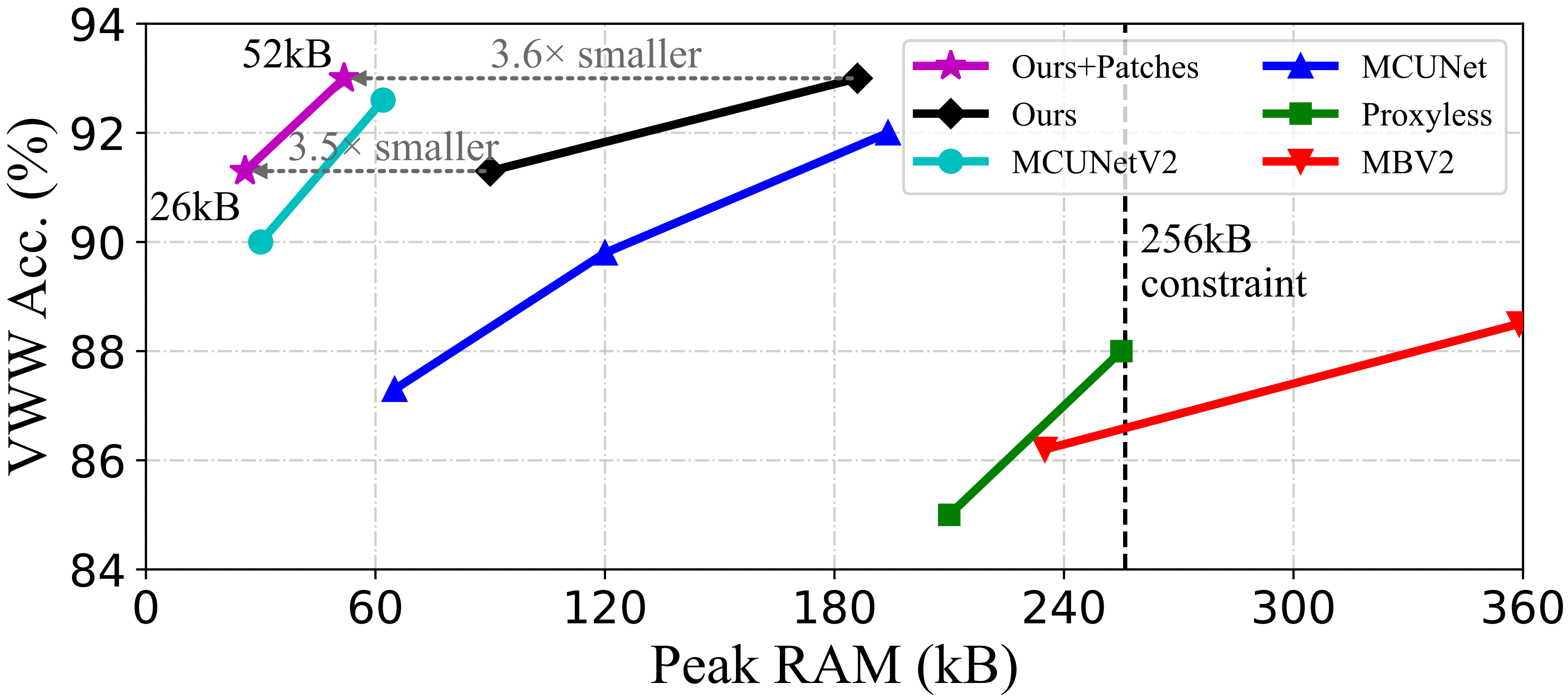
Table 4: Comparison of ImageNet classification accuracy on IoT devices

Model	Quant.	256kB SRAM, 1MB Flash			320kB SRAM, 1MB Flash			512kB SRAM, 2MB Flash		
		Mem	Size	Acc.	Mem	Size	Acc.	Mem	Size	Acc.
MBV1 [14, 26]	mixed	<256kB	<1MB	60.2%	-	-	-	<512kB	<2MB	68.0%
MBV2 [28]	8-bit	-	-	-	308kB	0.72MB	49.0%	-	-	-
Proxyless [3]	8-bit	-	-	-	292kB	0.72MB	56.2%	-	-	-
MCUNet-int8 [17]	8-bit	238kB	0.70MB	60.3%	293kB	0.70MB	61.8%	452kB	1.65MB	68.5%
MCUNet-int4 [17]	4-bit	233kB	0.67MB	62.0%	282kB	0.67MB	63.5%	498kB	1.56MB	70.7%
MCUNetV2 [16]	8-bit	196kB	0.79MB	64.9%	-	-	-	465kB	1.67MB	71.8%
Ours	mixed	253kB	0.73MB	**66.5%**	308kB	0.71MB	**68.2%**	507kB	1.67MB	**72.8%**

✓Under the tight constraints of 256kB SRAM and 1MB Flash, our model significantly improves the TOP-1 accuracy over quantized MCUNets.

✓Whole Table 4 indicates our QE-Score can specialize higher-capacity structures on resource-constrained IoT devices.

Experimental Results – Classification on VWW



✓Visual Wake Words (VWW) represents a realistic IoT use-case of identifying person.

✓Our model is superior to MCUNet in both accuracy and memory utilization.

Experimental Results – Detection on WIDER FACE

Alibaba Group 阿里巴巴集团

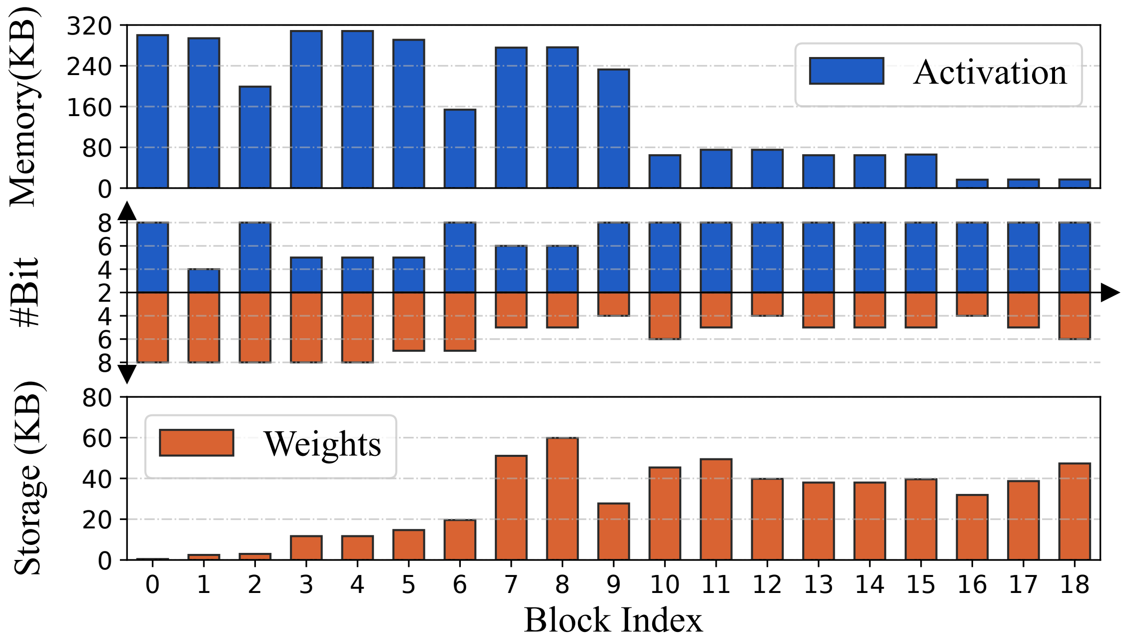
Table 6: Comparison of face detection on WIDER FACE. The hard subset is the most authoritative benchmark since it contains the faces in easy and medium subsets [19].

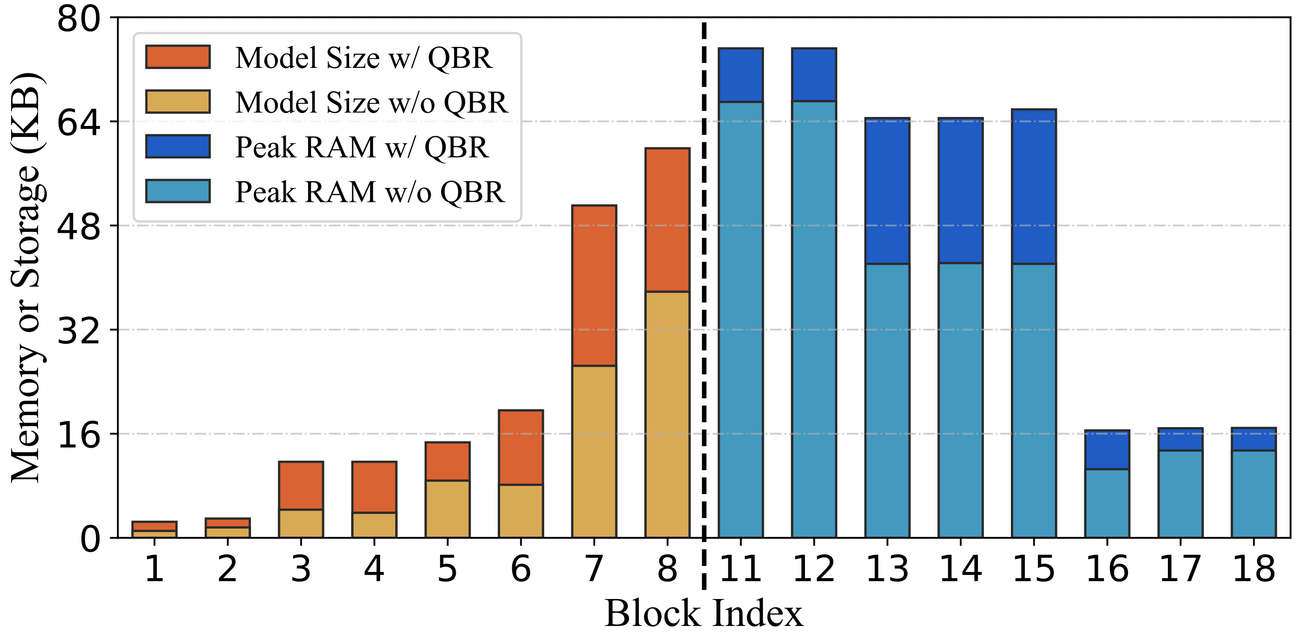
Model	Peak RAM	MACs	mAP		
			Easy	Medium	Hard
EagleEye [41]	1.17MB	0.08G	0.74	0.70	0.44
RNNPool [27]	1.17MB	0.10G	0.77	0.75	0.53
MCUNetV2 [16]	762kB	0.11G	**0.85**	0.81	0.55
Ours-Face	**650kB**	**0.04G**	0.82	**0.81**	**0.77**

✓To verify the generalization ability of our method, we conduct an experiment on Detection.

✓Our model can achieve a competitive mAP performance on WIDER Face dataset.

Experimental Results – QBR Visualization





✓Our method maintains higher-precision weights and lower-precision activations in the front few layers, while opposite in the latter few layers.

✓So QBR can strengthen resource utilization by embedding prior design knowledge.

Conclusion

✓To the best of our knowledge, we first present the ranking strategy of mixed-precision quantization networks in the entropy view to measure the expressiveness of the network.

✓Quantization Bits Refinement is proposed to adjust mixed quantization bits, which maximize the utilization of memory and storage resources on the IoT devices.

✓Benefitting from the QE-Score, our approach can achieve architecture searching within less than half a CPU hour.

Contact Xiuyu by DingTalk

Contact Xiuyu by WeChat

Thank you for your listening