CVPR 2021

Distribution-aware Adaptive Multi-bit Quantization

Sijie Zhao[1,2], Tao Yue[1], Xuemei Hu[1]

Nanjing University [1], Shenzhen Institute of Future Media Technology [2]

Problem Setting

- How to quantize the neural network parameters with lower precision and higher accuracy?
- How to allocate the bit-width to quantize different parts of weights and activations?

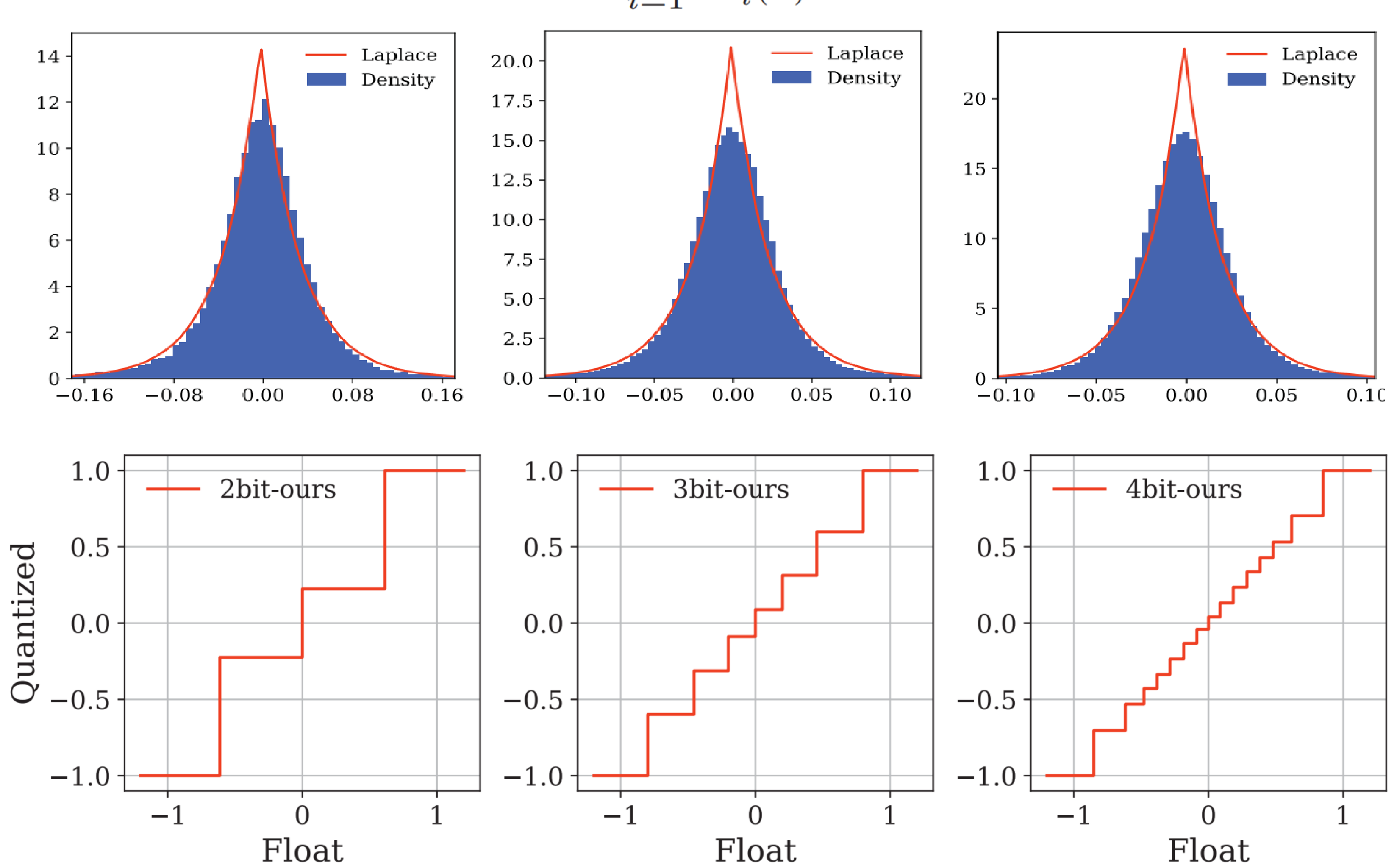
Contribution

- We introduce a distribution-aware multi-bit quantization (DMBQ) method for efficient and optimal MBQ quantization.
- We propose a first-order Taylor expansion based metric for evaluating the loss-sensitivity of the quantized weights and activations and introduce a loss-guided bit-width allocation (LBA) method.

Method

- We obtain the quantization scheme w.r.t different bit-width by minimize the expected multi-bit quantization error under a certain distribution.

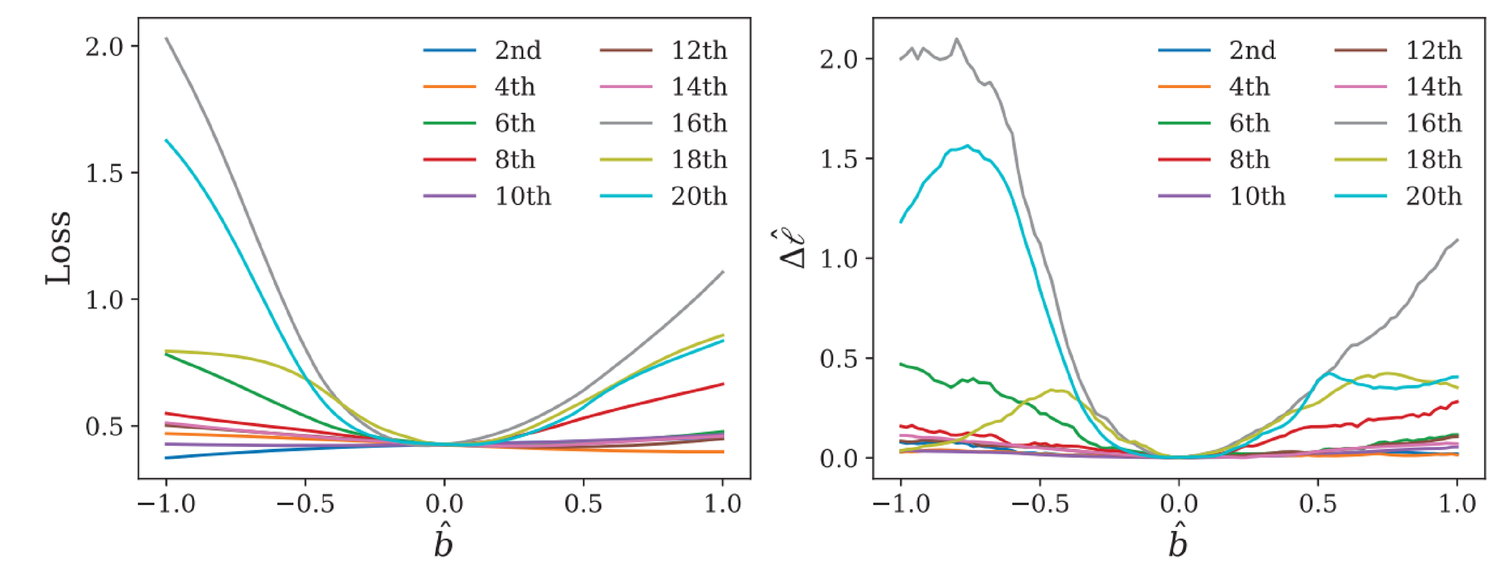
$$\min_{\boldsymbol{\alpha}} \mathbb{E}\left((X-\hat{X})^2\right) = \min_{\boldsymbol{\alpha}} \sum_{i=1}^{2^M} \int_{s_i(\boldsymbol{\alpha})}^{s_{i+1}(\boldsymbol{\alpha})} f(x)(x-q_i(\boldsymbol{\alpha}))^2 dx$$



Method

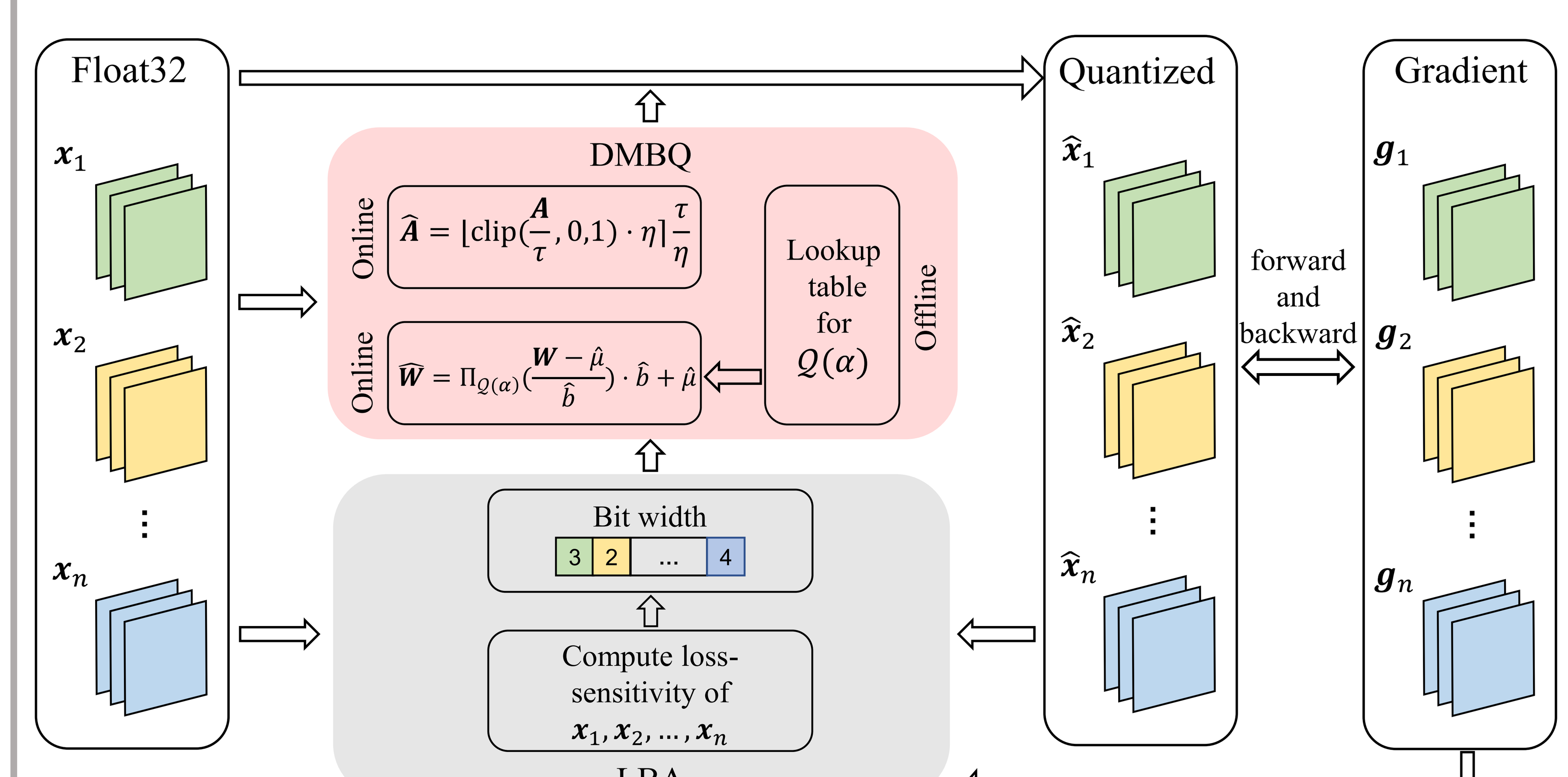
- We evaluate the quantization influence using Taylor expansion and quantize neural network into mixed-precision by gradients.

$$\Delta \ell' = \frac{|(\boldsymbol{x}-\hat{\boldsymbol{x}})^T \boldsymbol{g}(\hat{\boldsymbol{x}})|}{n}$$



Overview Framework

- The weights and activations are quantized by DMBQ in forward pass.
- The bit width is updated by LBA in backward pass where $\Delta l'$ is used as metric.



Experimental Results

Evaluation on ILSVRC12

Method	Prec (W/A)	Size(MB)	Top-1	Top-5
ResNet-18				
FP	32/32	46.7	70.3	89.5
APoT [18]	3/3	4.6	69.9	89.2
HAWQ [33]	-/-	6.1	68.6	-
AutoQ [22]	3.7/3.2	5.7	67.5	-
Ours	**3.0/3.0**	**4.7**	**70.0**	**89.4**
TTQ [40]*	2/32	4.9	66.6	87.2
INQ [38]	3/32	4.4	68.1	88.4
LQ-Net [36]*	2/32	4.9	68.0	88.0
ALQ [27]	2.0/32	3.4	68.9	-
Ours	**2.0/32**	**3.4**	**70.1**	**89.3**
BWN [28]*	1/32	3.5	60.8	83.0
HWGQ [2]*	1/32	3.5	61.3	-
DSQ [11]*	1/32	3.5	63.7	-
ALQ [27]	1.0/32	1.8	65.6	-
Ours	**1.0/32**	**1.8**	**65.9**	**87.1**
PACT [4]*	2/2	4.9	64.4	-
LQ-Net [36]*	2/2	4.9	64.9	85.9
DSQ [11]*	2/2	4.9	65.2	-
AutoQ [22]	2.2/3.0	3.6	66.4	-
ALQ [27]	2.0/2	3.4	66.4	-
Ours	**2.0/2.0**	**3.4**	**67.8**	**88.1**
PACT [4]*	1/2	3.5	62.9	-
LQ-Net [36]*	1/2	3.5	62.6	84.3
ALQ [27]	1.0/2	1.8	63.2	-
Ours	**1.0/2.0**	**1.8**	**63.5**	**85.5**
ResNet-34				
FP	32/32	87.1	73.7	91.3
LQ-Net [36]*	2/2	7.5	69.8	89.1
DSQ [11]*	2/2	7.4	70.0	-
ALQ [27]	2.0/2	6.3	71.1	-
Ours	**2.0/2.0**	**6.3**	**72.1**	**90.7**
HWGQ [2]*	1/2	4.8	64.3	85.7
LQ-Net [36]*	1/2	4.8	66.6	86.9
ALQ [27]	1.0/2	3.4	67.3	-
Ours	**1.0/2.0**	**3.4**	**69.8**	**89.2**

Evaluation on CIFAR10

Method	Prec (W/A)	Top-1
ResNet-20		
FP	32/32	92.4
LQ-Net [36]	2/32	91.8
Ours	**2.0/32**	**92.5**
BWN [28]	1/32	90.1
LQ-Net [36]	1/32	90.1
DSQ [11]	1/32	90.2
Ours	**1.0/32**	**91.4**
LQ-Net [36]	2/2	90.2
APoT [18]	2/2	91.0
Ours	**2.0/2.0**	**91.7**
LQ-Net [36]	1/2	88.4
Ours	**1.0/2.0**	**90.4**
VGG-small		
FP	32/32	93.8
BWN [28]	1/32	90.1
LQ-Net [36]	2/32	93.8
ALQ [27]	0.7/32	92.0
Ours	**0.7/32**	**93.7**
HWGQ [2]	1/2	92.5
LQ-Net [36]	1/2	93.4
Ours	**1.0/2.0**	**93.9**

Training Time

Method	Prec(W/A)	Time
FP	32/32	1.00×
LQ-Net [36]	2/32	1.40×
ALQ [27]	2.0/32	2.46×
DMBQ + LBA	**2.0/32**	**1.16×**
LQ-Net [36]	2/2	2.30×
LQ-Net [36]	3/3	3.70×
DMBQ	**4/4**	**1.14×**
DMBQ + LBA	**2.0/2.0**	**1.22×**

Ablation Study

Model	Method	Prec (W)	Top-1
VGG small	GP/LP	1	92.4
	CP	**0.7**	**93.7**
ResNet-18	GP	2	68.5
	LP	2.0	69.6
	CP	**2.0**	**70.1**
ResNet-18	GP/LP	1	64.5
	CP	**1.0**	65.9

Method	Prec (W/A)	Top-1	Top-5
FP	32	70.3	89.5
Uniform	2/2	62.7	84.6
	3/3	68.5	88.4
	4/4	70.0	89.3
DMBQ	2/2	65.1	86.4
	3/3	69.2	88.8
	4/4	70.2	89.4