THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY
MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE
Carnegie Mellon University
Meta

SDQ: Stochastic Differentiable Quantization with Mixed Precision

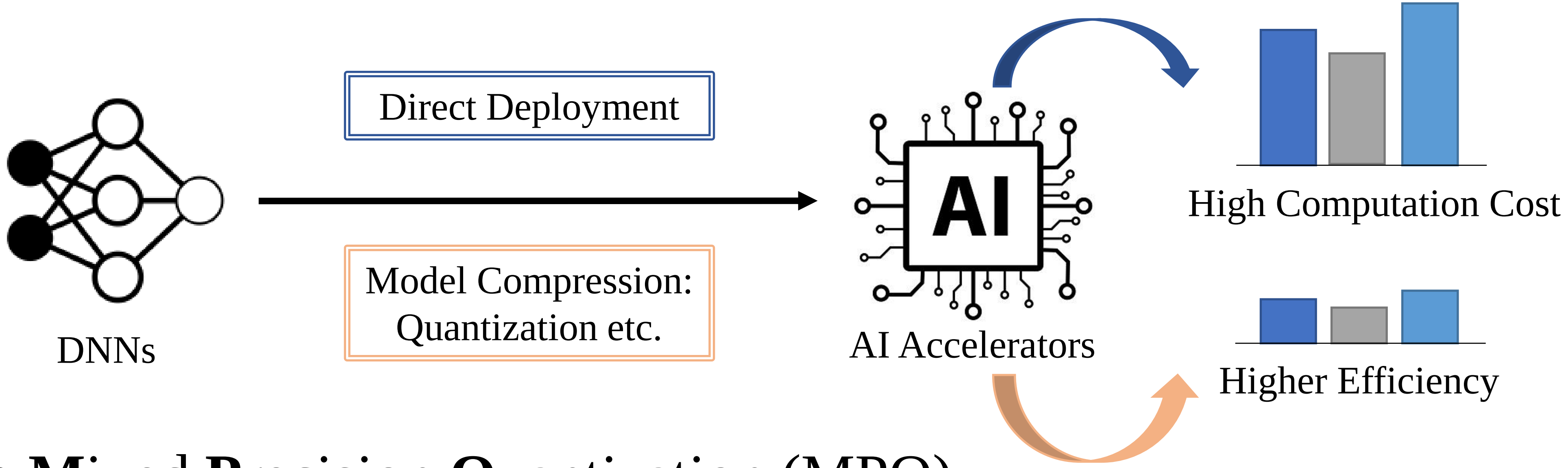
ICML 2022

Xijie Huang, Zhiqiang Shen, Shichao Li, Zechun Liu, Xianghong Hu,
Jeffry Wicaksana, Eric Xing, Kwang-Ting Cheng

Webpage: https://huangowen.github.io/SDQ/

Background

- Efficient deployment of deep learning models:
  - Consideration: time latency, energy consumption, model size



- **M**ixed **P**recision **Q**uantization (MPQ)
  - Fully leverage the various representation capacity for different modules

Motivation

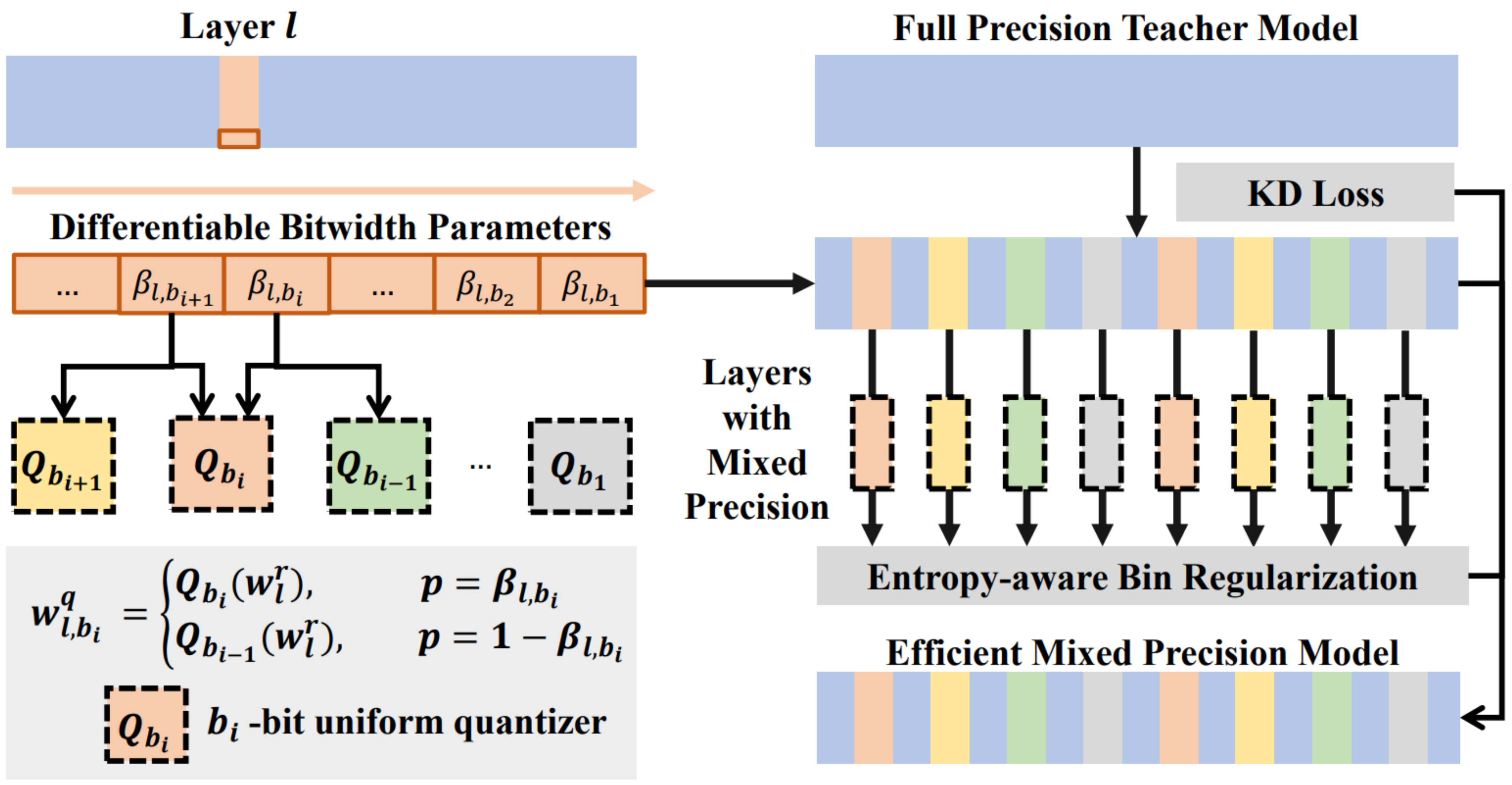
- Previous MPQ methods:
  - Search-based, Metric-based, Optimization-based

- Challenges:
  - Search-based: high computation cost of NAS or RL
  - Metric-based: sub-optimal generated MPQ strategy
  - Optimization-based: inaccurate gradient approximation

- Our target: fully differentiable, accurate, efficient method

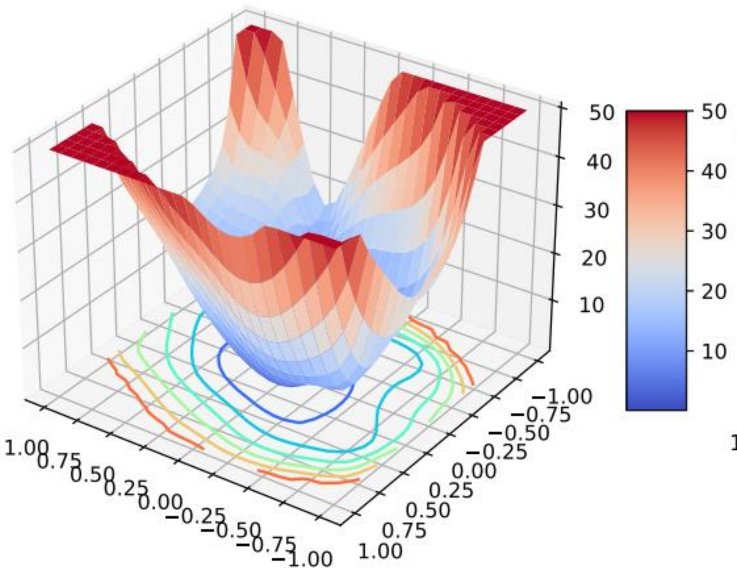
SDQ: **S**tochastic **D**ifferentiable **Q**uantization

- Learn the optimal bitwidth with **stochastic** quantizer

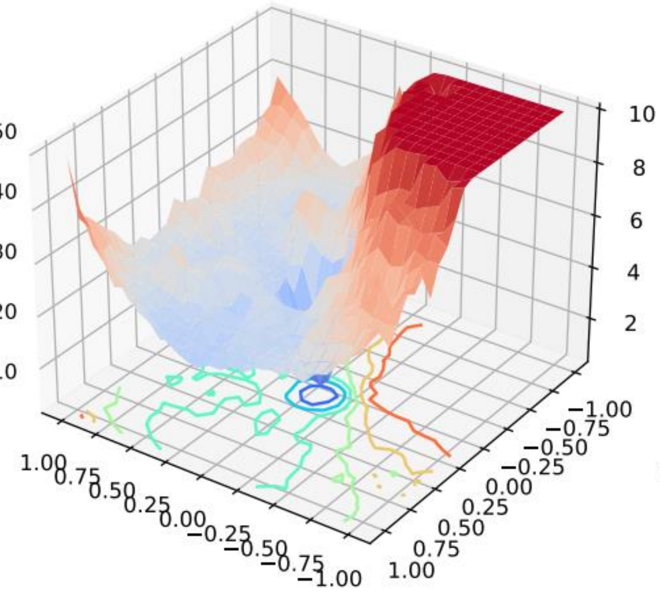


SDQ: **S**tochastic **D**ifferentiable **Q**uantization

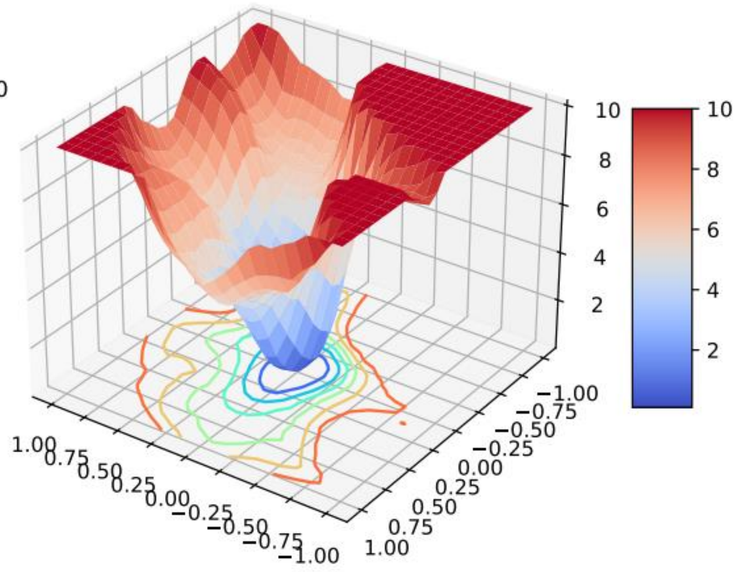
- **Differentiability**: Gumbel Softmax to sample from bitwidth
- Underlying loss optimization landscape:



Full Precision



Interpolation [Yang. L, AAAI’21]



Ours

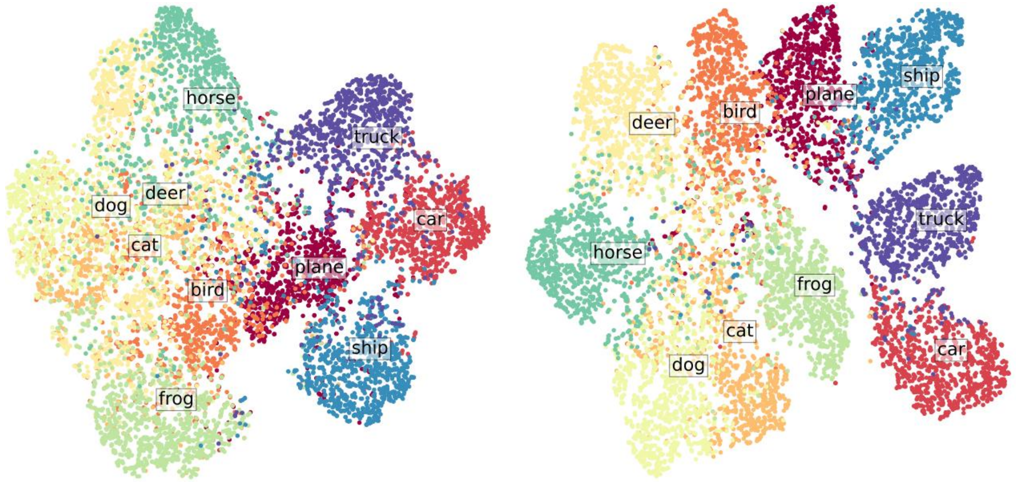
Experiment

- Networks: ResNet18, ResNet20, MobileNet-V2
- Datasets: CIFAR-10, ImageNet-1K

Comparison with SOTA of ResNet20 on CIFAR-10

Method	Bit-width (W/A)	mixed	Accuracy(%) Top-1	FP Top-1	WCR
Dorefa(Zhou et al., 2016)	2/32		88.2	92.4	16×
PACT (Choi et al., 2018)	2/32		89.7	92.4	16×
LQ-net(Zhang et al., 2018)	2/32		91.1	92.4	16×
TTQ (Jain et al., 2019)	2.00/32	✓	91.2	92.4	16×
Uhlich et al. (Uhlich et al., 2020)	2.00/32	✓	91.4	92.4	16×
BSQ (Yang et al., 2020)	2.08/32	✓	91.9	92.6	15.4×
DDQ (Zhang et al., 2021)	2.00/32	✓	91.6	92.4	16×
Ours	1.93/32	✓	**92.1**	92.4	**16.6×**

Feature embedding visualization using t-SNE



Dorefa (2-bit) Ours (1.93-bit)

Experiment

- Comparison with SOTA of ResNet18 on ImageNet-1K
  - Achieve an increase of 1.1% with a more compact bitwidth (3.61/4 vs. 4/4)

Network	Method	Bit-width (W/A)	Mixed	Uniform	Accuracy (%) Top-1	FP Top-1	WCR	Model Size (MB)	BitOPs (G)
ResNet18	Dorefa†(Zhou et al., 2016)	4/4		✓	68.1	70.5	8×	5.8	35.2
	PACT†(Choi et al., 2018)	4/4		✓	69.2	70.5	8×	5.8	35.2
	LQ-net (Zhang et al., 2018)	4/4			69.3	70.5	8×	5.8	35.2
	APOT (Li et al., 2019b)	4/4			70.7	70.5	8×	5.8	34.7
	DNAS†(Wu et al., 2018)	-/-	✓	✓	70.6	71.0	8×	5.8	35.2
	HAQ (Wang et al., 2019)	4/32	✓	✓	70.4	70.5	8×	5.8	465
	EdMIPS (Cai & Vasconcelos, 2020)	4/4	✓	✓	68.0	70.2	8×	5.8	34.7
	HAWQ-V3†(Yao et al., 2021)	4.8/7.5	✓	✓	70.4	71.5	6.7×	7.0	72.0
	Chen et al. (Chen et al., 2021)	3.85/4	✓	✓	$69.7_{\downarrow 0.1\%}$	69.8	8.3×	5.6	33.4
	FracBits-SAT (Yang & Jin, 2021)	4/4	✓	✓	$70.6_{\uparrow 0.4\%}$	70.2	8×	5.8	34.7
	Uhlich et al. (Uhlich et al., 2020)	3.88/4	✓		70.1	70.3	8.3×	5.6	33.7
	RMSMP (Chang et al., 2021)	4/4	✓		70.7	70.3	8×	5.8	34.7
	DDQ (Zhang et al., 2021)	4/4	✓		71.2	70.5	8×	5.8	34.7
	Ours	3.61/8	✓	✓	$\mathbf{72.1}_{\uparrow \mathbf{1.6\%}}$	70.5	**8.9**×	5.2	62.6
		3.61/4	✓	✓	$\mathbf{71.7}_{\uparrow \mathbf{1.2\%}}$	70.5	**8.9**×	5.2	**31.3**
		3.61/3	✓	✓	$\mathbf{70.2}_{\downarrow \mathbf{0.3\%}}$	70.5	**8.9**×	5.2	**23.5**
		3.61/2	✓	✓	$\mathbf{69.1}_{\downarrow \mathbf{1.4\%}}$	70.5	**8.9**×	5.2	**15.7**

Experiment

- Comparison with SOTA of MobileNetV2 on ImageNet-1K
  - Achieve a higher compression (8.7× vs. 8×) with higher accuracy (72.0% vs. 71.8%)
  - First model outperforms full-precision baseline with less than 4-bit settings

Network	Method	Bit-width (W/A)	Mixed	Uniform	Accuracy (%) Top-1	Accuracy (%) FP Top-1	WCR	Model Size (MB)	BitOPs (G)
MobileNetV2	Dorefa† (Zhou et al., 2016)	4/4		✓	61.8	71.9	8×	1.8	7.42
	PACT† (Choi et al., 2018)	4/4		✓	61.4	71.9	8×	1.8	7.42
	LQ-net (Zhang et al., 2018)	4/4			64.4	71.9	8×	1.8	7.42
	APOT (Li et al., 2019b)	4/4			71.0	71.9	8×	1.8	5.35
	HAQ (Wang et al., 2019)	4/32	✓	✓	71.5	71.9	8×	1.8	42.8
	HMQ (Habi et al., 2020)	3.98/4	✓	✓	70.9	71.9	8.1×	1.7	5.32
	Chen et al. (Chen et al., 2021)	4.27/8	✓	✓	$71.8_{\downarrow 0.1\%}$	71.9	7.5×	1.9	5.32
	FracBits-SAT (Yang & Jin, 2021)	4/4	✓	✓	$71.6_{\downarrow 0.2\%}$	71.8	8×	1.8	5.35
	Uhlich et al. (Uhlich et al., 2020)	3.75/4	✓		69.8	70.2	8.5×	1.6	5.01
	RMSMP (Chang et al., 2021)	4/4	✓		69.0	71.9	8×	1.8	5.35
	DDQ (Zhang et al., 2021)	4/4	✓		71.8	71.9	8×	1.8	5.35
	Ours	3.66/8	✓	✓	$\mathbf{72.9}_{\uparrow 1.0\%}$	71.9	**8.7**×	1.8	9.79
		3.66/4	✓	✓	$\mathbf{72.0}_{\uparrow 0.1\%}$	71.9	**8.7**×	1.8	**4.89**

Conclusion

- We present a novel stochastic quantization framework to learn the optimal mixed precision quantization strategy.
- We utilize the straight-through Gumbel-Softmax estimator in the gradient computation w.r.t. differentiable bitwidth parameters.
- We extensively evaluate our method on different networks (ResNet18 and MobileNetV2) and datasets (CIFAR-10 and ImageNet-1K).
- More in the paper:
  - Entropy-aware Bin Regularizer (EBR) to minimize quantization error
  - Knowledge distillation loss and analysis on different full-precision teacher
  - Deployment experiments on the object detection task and a real FPGA system

ICML
International Conference
On Machine Learning

Thanks for listening!

Webpage: https://huangowen.github.io/SDQ/