### QReg: On Regularization Effects of Quantization

MohammadHossein AskariHemmat<sup>1</sup>, Reyhane Askari Hemmat<sup>2</sup>, Alex Hoffman<sup>3</sup>, Ivan Lazarevich<sup>3</sup>, Sudhakar Sah<sup>3</sup>, Ehsan Saboori<sup>3</sup>, Olivier Mastropietro<sup>3</sup>, Yvon Savaria<sup>1</sup>, Jean-Pierre David<sup>1</sup> <sup>1</sup>Ecole Polytechnique Montreal, <sup>2</sup>Mila, Universite de Montreal, <sup>3</sup>Deeplite Inc. Montreal

> ICML 2022 HAET Workshop July 23 2022







### **Outline:**

- **Related Works** ullet
- Modeling Quantization as an Additive Noise
- Experiments and Results
- Conclusion ullet
- Acknowledgement





### **Related Works:**

Regularization effect of quantization has been studied before:

- 1. Effect of Quantization on Accuracy Improvement.
- 2. Analytical Studies.
- 3. Using Quantization for its Regularization Effect.





# Related Works (2):

Regularization effect of quantization has been studied before:

- 1. Effect of Quantization on Accuracy Improvement.
- 2. Analytical Studies.
- 3. Using Quantization for its Regularization Effect.

Our Contribution:

- 1. Studying relationship between quantization level and regularization effect.
- datasets and tasks.



POLYTECHNIQUE

2. Providing empirical study over different quantization levels and methods and different models,



### Modeling Quantization as an Additive Noise (1):

Weight quantization can be modeled as weight perturbation:

### $f(x, w_q)$ : Is the predicted target of the network f parameterized by w $\delta$



$$= f(x, w + \delta)$$

We assumed, quantization noise follows a normal distribution:  $\delta \sim \mathcal{N}(0, \sigma I)$ 





## Modeling Quantization as an Additive Noise (2):

For simplicity, we consider a regression problem:

$$\mathcal{L} = \frac{1}{m} \sum_{i=1}^{m} \|\hat{y}_i - y_i\|_2^2$$

### Applying a first-order Taylor approximation around the weights of the full precision model:

 $\widetilde{\mathcal{L}} \approx \mathcal{L} + \frac{q}{2}$ 



$$\frac{\sigma \delta^2}{m} \sum_{i=1}^m \|\nabla w \hat{y_i}\|_2^2$$





## Modeling Quantization as an Additive Noise (3):

For simplicity, we consider a regression problem:







NIVERSITÉ 'INGÉNIERIE





## **Experiments and Results (1):**

We tested regularization effect of quantization over different:

- 1. Models:
  - Resnet18, Resnet20, Resnet50, Mobilenet V1, Yolo5n.
- Datasets: 2.
  - CIFAR10, CIFAR100, VOC.
- Quantization methods: 3.
  - LSQ, PACT, DoReFa.
- Quantization levels: 4.
  - 2 bits, 4 bits, 8 bits, FP32



MONTRÉAL

POLYTECHNIQUE



## Experiments and Results (2):

For each test, we used different augmentations on original dataset:

Test Set	Brightness	Contrast	Defocus Blur	Elastic Transform
Fog	Frost	Gaussian Blur	Gaussian Noise	Glass Blur
Impulse Noise	Jpeg Compression	Motion Blur	Pixelate	Saturate
Shot Noise	Snow	Spatter	Speckle Noise	Zoom Blur





UNIVERSITÉ D'INGÉNIERIE



### **Experiments and Results (3):**

The value in each cell corresponds to difference between quantized and FP32 model:

3.60	4.20	3.40	B.80	3.50
3.90	3.10	3.70	-0.40	0.20
2.80	2.60	2.00	3.80	2.20
1.30	3.60	3.10	1.10	2.50





UNIVERSITÉ D'INGÉNIERIE



## **Experiments and Results (3):**

The value in each cell corresponds to difference between quantized and FP32 model:

		3.50
		0.20





JNIVERSITÉ D'INGÉNIERIE

**Quantized model performed better** 

### QReg: Regularization Effect of Quantization



11

## **Experiments and Results (3):**

The value in each cell corresponds to difference between quantized and FP32 model:





NIVERSITÉ D'INGÉNIERIE





### **Experiments and Results (4) LSQ Quantization:** 4 bits 2 bits 8 bits

1. Nine different experiments.

Resnet20







JNIVERSITÉ D'INGÉNIERIE

				-7.8	-0.2	0.3	-0.2	0.5	0.7	0.7	1.1	0.7	1.7	1
				27.1	-0.1	1.1	0.7	0.7	7.0	0.7	1.2	1.6	-0.1	2
-1.9	-1.4			-9.2	2.9	0.7	0.0	0.9	-0.1	0.3	1.1	1.9	1.3	0
-0.2	-4.4	-6.3	-0.4	-6.3	1.3	2.2	2.2	0.7	0.2	1.1	1.9	2.1	1.2	2
-0.6	0.0	-0.5	-0.7	-0.3	2.7	3.1	2.3	2.4	1.2	3.6	4.2	3.4	3.8	3
-0.3	-0.7	-0.6	-4.4	0.3	2.6	1.3	2.7	-2.2	-1.5	3.9	3.1	3.7	-0.4	0
1.9	-0.7	-0.3	-0.6	-2.8	3.8	1.7	1.0	1.4	1.7	2.8	2.6	2.0	3.8	2
-2.7	-0.6	-0.8	-3.2	-1.5	-0.3	2.0	2.4	-0.2	0.4	1.3	3.6	3.1	1.1	2
-17.8	-17.4	-13.4	-12.0	-12.3	-4.2	-4.9	-4.3	-5.0	-4.2	3.3	3.2	-1.0	1.0	0
				-15.7	-4.4	-1.3		-2.4	-5.4	-0.4	2.7	2.1	2.0	2
			-10.8	-17.4	-2.3	-4.6		-4.0	-4.8	1.3	2.9	1.4	-0.2	2
				-7.1	-2.6	-1.0	-4.7	-3.5	-3.1	2.1	3.1	3.2	1.6	-0

QReg: Regularization Effect of Quantization









### **Experiments and Results (4) LSQ Quantization:** 2 bits 4 bits 8 bits

- 1. Nine different experiments.
- 2. Three different quantization levels.

Sn

Resnet20





NIVERSITÉ INGÉNIERIE/

				-7.8	-0.2	0.3	-0.2	0.5	0.7	0.7	1.1	0.7	1.7	1
				27.1	-0.1	1.1	0.7	0.7	7.0	0.7	1.2	1.6	-0.1	2
-1.9	-1.4			-9.2	2.9	0.7	0.0	0.9	-0.1	0.3	1.1	1.9	1.3	0
-0.2	-4.4		-0,4	-6.3	1.3	2.2	2.2	0.7	0.2	1.1	1.9	2.1	1.2	2
-0.6	0.0	-0.5	-0.7	-0.3	2.7	3.1	2.3	2.4	1.2	3.6	4.2	3.4	3.8	3
-0.3	-0.7	-0.6	-4.4	0.3	2.6	1.3	2.7	-2.2	-1.5	3.9	3.1	3.7	-0.4	0
1.9	-0.7	-0.3	-0.6	-2.8	3.8	1.7	1.0	1.4	1.7	2.8	2.6	2.0	3.8	2
-2.7	-0.6	-0.8	-3.2	-1.5	-0.3	2.0	2.4	-0.2	0.4	1.3	3.6	3.1	1.1	2
				-12.3	-4.2	-4.9	-4.3		-4.2	3.3	3.2	-1.0	1.0	0
				-15.7	-4.4	-1.3		-2.4	-5.4	-0.4	2.7	2.1	2.0	2
			-10.8	-17.4	-2.3	-4.6		-4.0	-4.8	1.3	2.9	1.4	-0.2	2
-15.3	-8.8	-17.5	-15.6	-7.1	-2.6	-1.0	-4.7	-3.5	-3.1	2.1	3.1	3.2	1.6	-(









### Experiments and Results (4) LSQ Quantization: 2 bits 4 bits 8 bits

Resnet20

 $\infty$ 

Resnet1

E S

- 1. Nine different experiments.
- 2. Three different quantization levels.
- Three different models: 3.
  - 1. Resnet20: CIFAR10
  - 2. Resnet18: CIFAR100
  - 3. YOLOv5n: VOC





-9.0	-9.0	-10.5	-8.4	-7.8	-0.2	0.3	-0.2	0.5	0.7	0.7	1.1	0.7	1.7	1
					-0.1	1.1	0.7	0.7		0.7	1.2	1.6	-0.1	2
-1.9	-1.4				2.9	0.7	0.0	0.9	-0.1	0.3	1.1	1.9	1.3	0
-0.2	-4.4	-6.3	-0.4	-6.3	1.3	2.2	2.2	0.7	0.2	1.1	1.9	2.1	1.2	2
-0.6	0.0	-0.5	-0.7	-0.3	2.7	3.1	2.3	2.4	1.2	3.6	4.2	3.4	3.8	3
-0.3	-0.7	-0.6	-4.4	0.3	2.6	1.3	2.7	-2.2	-1.5	3.9	3.1	3.7	-0.4	0
1.9	-0.7	-0.3	-0.6	-2.8	3.8	1.7	1.0	1.4	1.7	2.8	2.6	2.0	3.8	2
-2.7	-0.6	-0.8	-3.2	-1.5	-0.3	2.0	2.4	-0.2	0.4	1.3	3.6	3.1	1.1	2
-17.8	-17.4	-13.4	-12.0	-12.3	-4.2	-4.9	-4.3	-5.0	-4.2	3.3	3.2	-1.0	1.0	0
					-4.4	-1.3		-2.4		-0.4	2.7	2.1	2.0	2
			-10.8		-2.3	-4.6		-4.0	-4.8	1.3	2.9	1.4	-0.2	2
-15.3	-8.8	-17.5	-15.6	-7.1	-2.6	-1.0	-4.7	-3.5	-3.1	2.1	3.1	3.2	1.6	-(









### **Experiments and Resul**

- 1. Nine different experiments.
- 2. Three different quantization levels.
- Three different models: 3.
  - 1. Resnet20: CIFAR10
  - 2. Resnet18: CIFAR100
  - 3. YOLOv5n: VOC
- 4. 8-bit models perform consistently better.



ts	5	(4)	) [	_S	<b>S</b> Q	) (	Q	SL	n	tiz	za	ti	<b>Oľ</b>	1
		2 bit	ts				4 b	its			8	3 bit	ts	
-9.0	-9.0	-10.5	-8.4	-7.8	-0.2	0.3	-0.2	0.5	0.7	0.7	1.1	0.7	1.7	1
-10.2	-5.8	-8.6	6.4	27.1	-0.1	1.1	0.7	0.7	7.0	0.7	1.2	1.6	-0.1	2
-1.9	-1.4	-5.8	-6.3	-9.2	2.9	0.7	0.0	0.9	-0.1	0.3	1.1	1.9	1.3	0
-0.2	-4.4	-6.3	-0.4	-6.3	1.3	2.2	2.2	0.7	0.2	1.1	1.9	2.1	1.2	2
-0.6	0.0	-0.5	-0.7	-0.3	2.7	3.1	2.3	2.4	1.2	3.6	4.2	3.4	3.8	3
-0.3	-0.7	-0.6	-4.4	0.3	2.6	1.3	2.7	-2.2	-1.5	3.9	3.1	3.7	-0.4	0
1.9	-0.7	-0.3	-0.6	-2.8	3.8	1.7	1.0	1.4	1.7	2.8	2.6	2.0	3.8	2
-2.7	-0.6	-0.8	-3.2	-1.5	-0.3	2.0	2.4	-0.2	0.4	1.3	3.6	3.1	1.1	2
-17.8	-17.4	-13.4	-12.0	-12.3	-4.2	-4.9	-4.3	-5.0	-4.2	3.3	3.2	-1.0	1.0	0
-13.3	-12.7	-17.1	-15.0	-15.7	-4.4	-1.3	-5.9	-2.4	-5.4	-0.4	2.7	2.1	2.0	2
-14.0	-17.6	-15.3	-10.8	-17.4	-2.3	-4.6	-5.4	-4.0	-4.8	1.3	2.9	1.4	-0.2	2
-15.3	-8.8	-17.5	-15.6	-7.1	-2.6	-1.0	-4.7	-3.5	-3.1	2.1	3.1	3.2	1.6	-0







### Experiments and Results (5) Relative Improvement:

1. Relative improvement score:

Error Improvement= lo





JNIVERSITÉ D'INGÉNIERIE

$$g(\frac{100 - fval}{100 - qval}) * 100$$

17

### **Experiments and Results (5) Relative Improvement:**

- 1. Relative improvement score.
- 2. 4-bit and 8-bit models have better generalization.

	Quantization	Avg Accuracy	Relative Error
	Level	Augmented Data	Improvement Eq.5
	2 bits	79.17	-8.32
Resnet20	4 bits	83.94	2.98
Cifar10	8 bits	84.07	3.33
	FP32	82.80	0.00
	2 bits	49.43	-0.84
Resnet18	4 bits	51.76	1.21
Cifar100	8 bits	53.05	2.39
	FP32	50.40	0.00
	2 bits	14.66	-7.86
YOLOv5n	4 bits	24.90	-2.31
VOC	8 bits	30.34	0.96
	FP32	28.78	0.00





NIVERSITÉ INGÉNIERIE/

18

### **Conclusion:**

- 1. We formalized quantization noise and study how it effects training.
- 2. We showed how quantization level is correlated to the regularization term.
- 3. We provided a extensive list of experiments where we tested our hypothesis on different models, tasks, quantization methods and levels.
- 4. Based on our study, we propose 8-bit quantization provides a reliable form of regularization in different vision tasks and models.





### Acknowledgements

The authors would like to thank:

- 1. Mohammad Pezeshki and Anush Sankaran for helpful discussion about designing empirical studies presented in this paper
- 2. FRQNT and NSERC for providing financial support for this project.





NIVERSITÉ D'INGÉNIERIE

