ZeroQuant-Series: Towards LLM Post-Training Quantization

Zhewei Yao

DeepSpeed of Microsoft

Outline

- ZeroQuant: Efficient and Affordable Post-Training Quantization for Large-Scale Transformers
- ZeroQuant-V2: Exploring Post-training Quantization in LLMs from Comprehensive Study to Low Rank Compensation
- ZeroQuant-FP: A Leap Forward in LLMs Post-Training W4A8 Quantization Using Floating-Point Formats

[1] [2206.01861] ZeroQuant: Efficient and Affordable Post-Training Quantization for Large-Scale Transformers (arxiv.org)
[2] https://arxiv.org/pdf/2303.08302.pdf
[3] https://arxiv.org/pdf/2307.09782.pdf

Outline

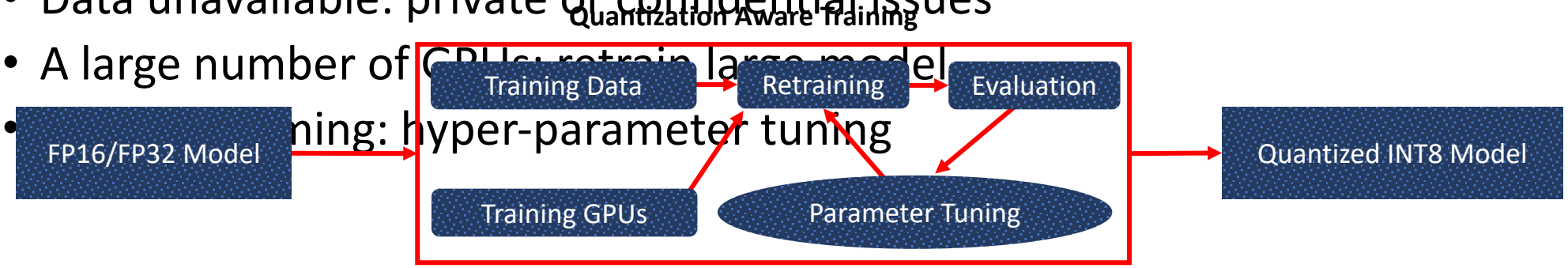
- ZeroQuant: Efficient and Affordable Post-Training Quantization for Large-Scale Transformers
- ZeroQuant-V2: Exploring Post-training Quantization in LLMs from Comprehensive Study to Low Rank Compensation
- ZeroQuant-FP: A Leap Forward in LLMs Post-Training W4A8 Quantization Using Floating-Point Formats

[1] [2206.01861] ZeroQuant: Efficient and Affordable Post-Training Quantization for Large-Scale Transformers (arxiv.org)
[2] https://arxiv.org/pdf/2303.08302.pdf
[3] https://arxiv.org/pdf/2307.09782.pdf

Challenges of inferencing large scale models

- Two main challenges of inferencing large scale models
  - High memory consumption: 40G A-100 for a ~20B model (FP16)
  - Slow speed: ~30ms for a token using GPT-J (6B)

- Quantization is one promising approach, but QAT …
  - Data unavailable: private or confidential issues
  - A large number of GPUs: retrain large model
  - …ning: hyper-parameter tuning



Adv. and Disadv. of post-training quantization

- PTQ has better compression efficiency
  - Portion of training data
  - A small amount of GPUs
  - Little to no retraining
- Directly applying PTQ leads to accuracy loss

BERT-Base GLUE Performance with various precisions

Precision	CoLA	MNLI-m	MNLI-mm	MRPC	QNLI	QQP	RTE	SST-2	STS-B	Ave.
W16A16	59.72	84.94	85.06	86.27/90.57	92.15	91.51/88.56	72.20	93.23	90.06/89.59	83.95
W8A16	60.77	84.65	84.92	85.29/89.86	91.84	91.52/88.56	71.84	93.46	89.89/89.50	83.87
W16A8	56.85	80.55	81.48	84.07/89.33	91.34	91.30/88.07	68.59	93.46	88.74/88.74	81.93
W8A8	58.74	79.99	81.06	84.31/89.51	91.18	91.24/88.03	70.76	92.66	88.33/88.73	82.16
W4/8A16	0.00	16.74	16.95	31.62/0.00	50.74	63.18/0.00	47.29	70.64	16.48/15.91	33.11

Adv. and Disadv. of post-training quantization

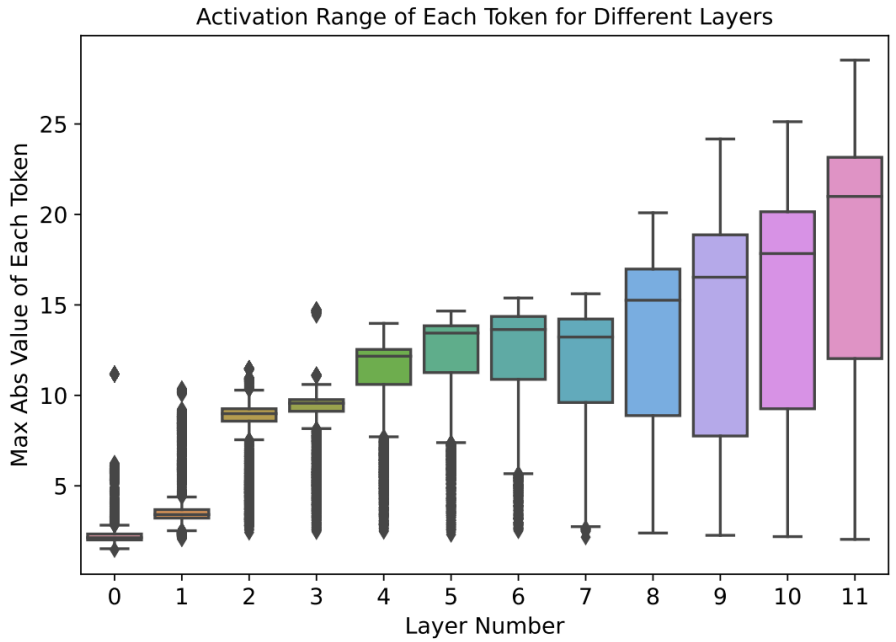
- PTQ has better compression efficiency
  - Portion of training data
  - A small amount of GPUs
  - Little to no retraining
- Directly applying PTQ leads to accuracy loss

Zero-shot evaluation of GPT-3-350M with various precisions

Precision	Lambada (↑)	PIQA (↑)	OpenBookQA (↑)	RTE (↑)	ReCoRd (↑)	Ave. 19 Tasks (↑)	Wikitext-2 (↓)
W16A16	49.3	66.3	29.4	53.8	75.1	38.9	21.5
W8A16	49.3	66.1	29.6	54.2	74.8	38.5	22.1
W16A8	44.7	64.8	28.2	52.7	69.2	37.8	24.6
W8A8	42.6	64.1	28.0	53.1	67.5	37.8	26.2
W4/8A16	0.00	51.4	30.2	52.7	16.1	28.9	1.76e5

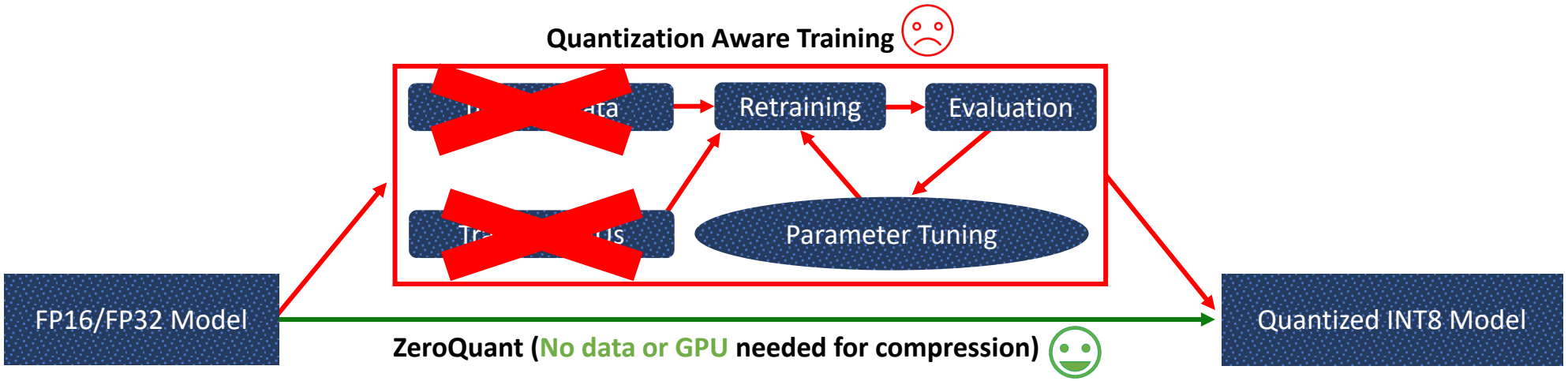
Why PTQ does not work

- Dynamic activation range
  - Different tokens have dramatically different activation ranges
- Different ranges of neurons in weight
  - No enough precision left for small-range neurons



ZeroQuant: INT8 without Compression Cost

- Quantize large-scale models within limited time/resource budget

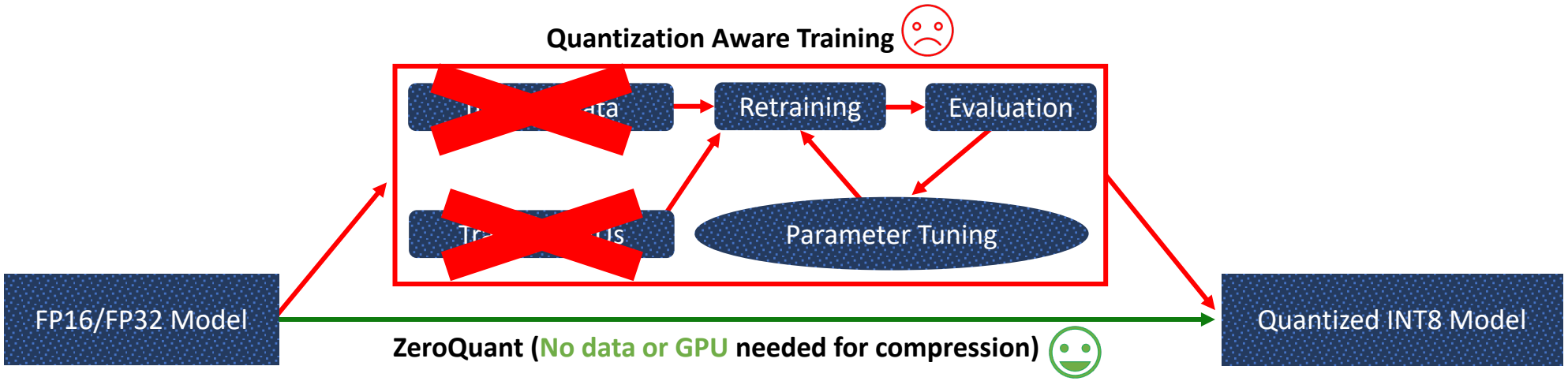


- Fine-grained quantization schemes to reduce quantization error

- Specified INT8 kernels to get real latency reduction

ZeroQuant: INT8 without Compression Cost

- Quantize large-scale models within limited time/resource budget



FP16 Weight Matrix

1.1	2.2	3.3	4.4	5.5	6.6
…					
…					
…					
…					
1.2	2.4	3.6	4.8	6.0	7.2

≈

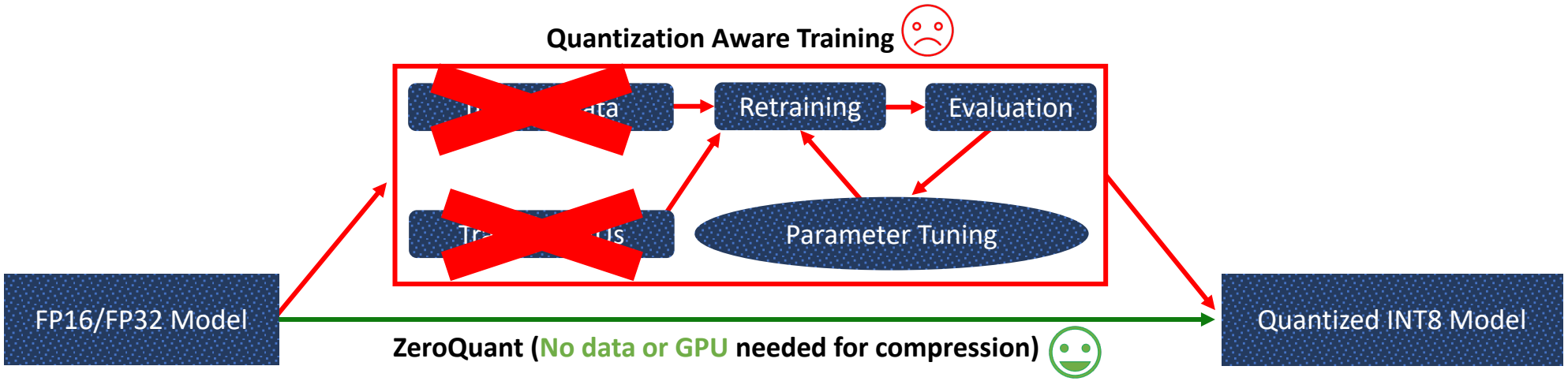
Scaling Factor: 1.1 *

QAT INT8

1	2	3	4	5	6
…					
…					
…					
…					
1	2	3	4	5	6

ZeroQuant: INT8 without Compression Cost

- Quantize large-scale models within limited time/resource budget



FP16 Weight Matrix

1.1	2.2	3.3	4.4	5.5	6.6
…					
…					
…					
…					
1.2	2.4	3.6	4.8	6.0	7.2

≈

ZeroQuant INT8

Scaling Factor: 1.1 *

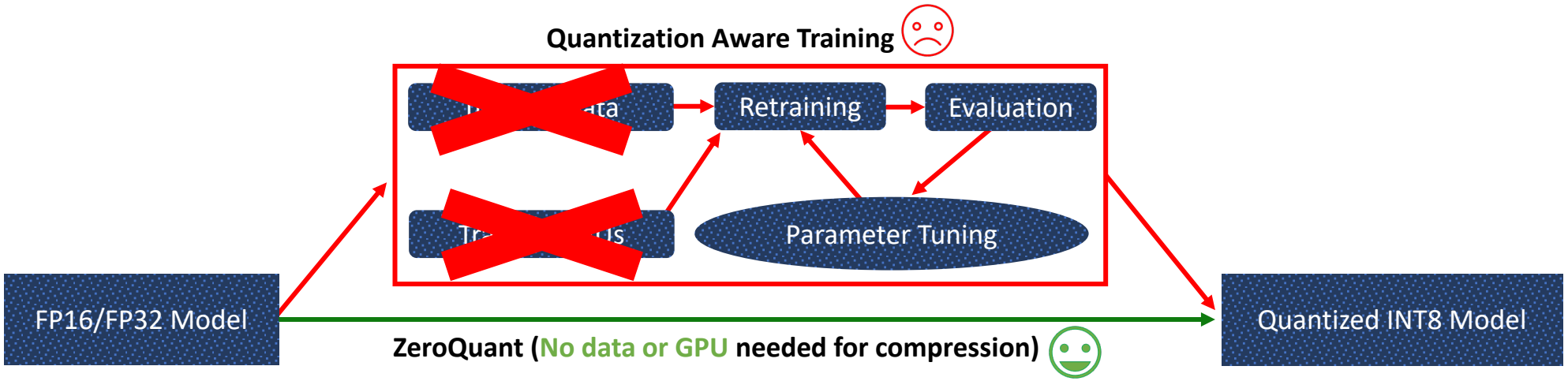
1	2	3	4	5	6
…					
…					

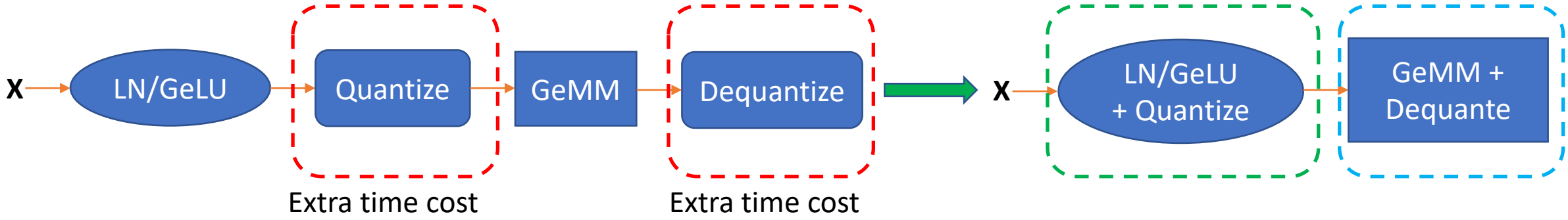
Scaling Factor: 1.2 *

…					
…					
1	2	3	4	5	6

ZeroQuant: INT8 without Compression Cost

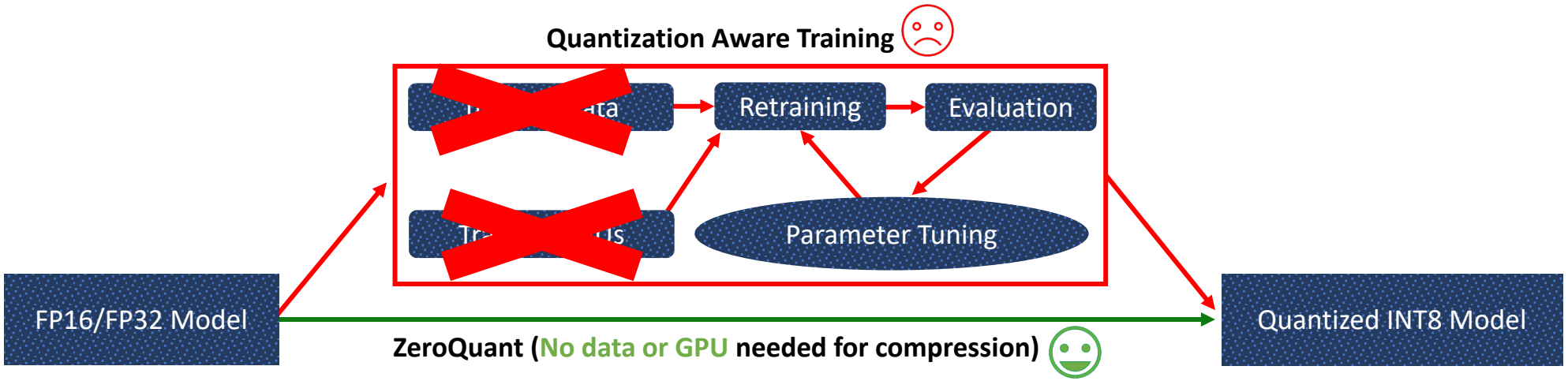
- Quantize large-scale models within limited time/resource budget





ZeroQuant: INT8 without Compression Cost

- Quantize large-scale models within limited time/resource budget

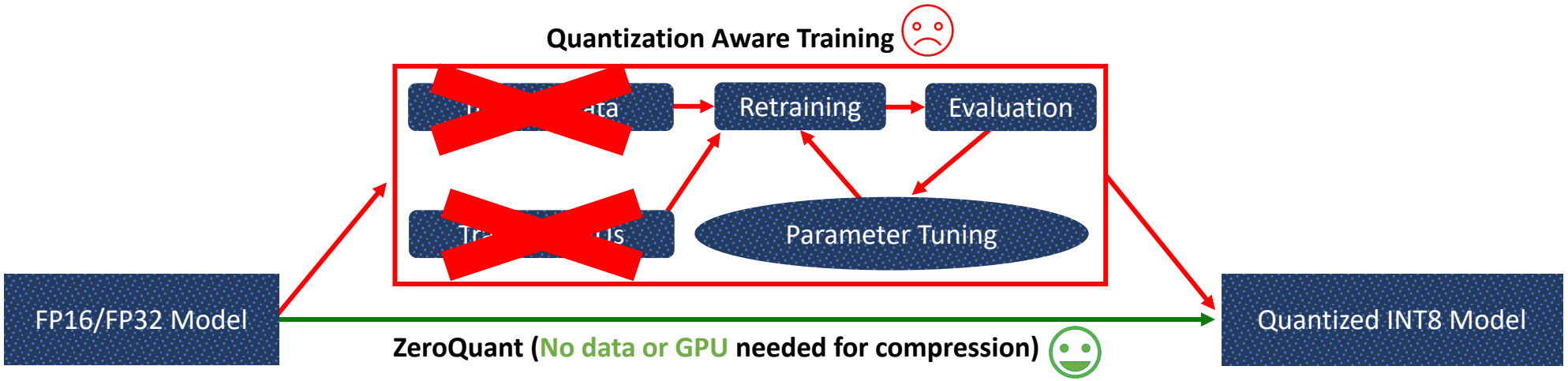


BERT-Base GLUE Performance with QAT and PTQ

Precision (Method)	CoLA	MNLI-m	MNLI-mm	MRPC	QNLI	QQP	RTE	SST-2	STS-B	Ave.	Ave. Time (s)
W16A16 (Baseline)	59.72	84.94	85.06	86.27/90.57	92.15	91.51/88.56	72.20	93.23	90.06/89.59	83.95	N/A
W8A8 [56] (QAT)$^+$	—	83.91	83.83	—	—	—	—	92.83	—	—	—
W8A8 [76] (QAT)	58.48	—	—	—/89.56	90.62	—/87.96	68.78	92.24	89.04/—	—	—
W8A8 (QAT)	61.21	84.80	84.64	83.82/88.85	91.29	91.29/88.28	71.12	92.89	88.39/88.18	83.37	2900
W8A8 (PTQ)	56.06	79.99	81.06	75.49/79.67	87.35	89.92/86.82	48.38	91.40	86.58/86.44	77.41	6
W8A8/16 [6] (PTQ)*	58.63	82.67	82.67	88.74	90.41	89.40	68.95	92.66	88.00	82.46	Unknown
W8A8 (ZeroQuant)	59.59	84.83	85.13	86.03/90.39	91.98	91.45/88.46	71.12	93.12	90.09/89.62	83.75	0

ZeroQuant: INT8 without Compression Cost

- Quantize large-scale models within limited time/resource budget

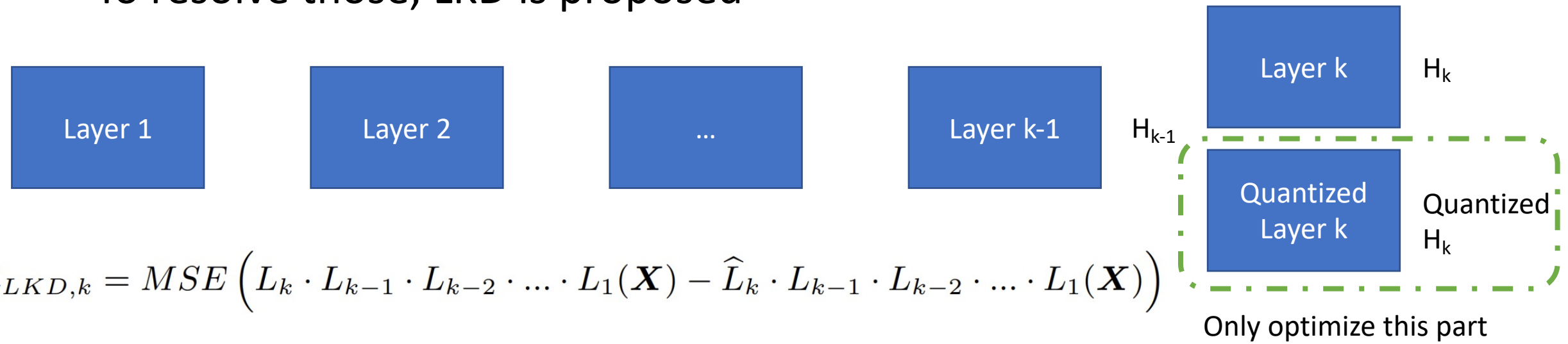


GPT-3-Style 125M	Ave. over 19 Tasks	Wikitext (lower better)	# GPUs for Compression	Time for Compression	Data Requirement
Baseline, FP16	36.31	29.4	N/A	N/A	N/A
QAT, INT8	35.99	33.24	32	20 hours	Yes
ZeroQuant, INT8	36.32	29.5	**None**	**None**	**No**

GPT-Neox 20B	LAMBADA	PIQA	Ave. over 19 Tasks	# GPUs for Compression	Time for Compression	Data Requirement	GPU x Latency (ms)	Inf Cost Reduction
Baseline, FP16	71.7	77.7	50.44				2x65	1x
QAT, INT8	--	--	--	96	20 days	Yes	--	--
ZeroQuant, INT8	71.9	78.3	50.38	**None**	**None**	**No**	**1x25**	**5.2x**

Layer-by-layer knowledge distillation

- Knowledge distillation for even lower-bit quantization
  - Hold a teacher and a student model together
  - Several copies (gradient, first/second order momentum) of the weight
  - Original training data
- To resolve those, LKD is proposed



$$\mathcal{L}_{LKD,k} = MSE\left(L_k \cdot L_{k-1} \cdot L_{k-2} \cdot \ldots \cdot L_1(\boldsymbol{X}) - \widehat{L}_k \cdot L_{k-1} \cdot L_{k-2} \cdot \ldots \cdot L_1(\boldsymbol{X})\right)$$

Layer-by-layer knowledge distillation

- Knowledge distillation for even lower-bit quantization
  - Hold a teacher and a student model together
  - Several copies (gradient, first/second order momentum) of the weight
  - Original training data
- To resolve those, LKD is proposed
  - No need a separate teacher
  - Reduced memory overhead
  - Work even without original training data

Mixed-precision results with LKD

- Without tuning, LKD
  - ~1 point gain with 31s on BERT-base

BERT-Base GLUE Performance with QAT and PTQ

Precision (Method)	CoLA	MNLI-m	MNLI-mm	MRPC	QNLI	QQP	RTE	SST-2	STS-B	Ave.	Ave. Time (s)
W16A16 (Baseline)	59.72	84.94	85.06	86.27/90.57	92.15	91.51/88.56	72.20	93.23	90.06/89.59	83.95	N/A
W8A8 [56] (QAT)$^+$	—	83.91	83.83	—	—	—	—	92.83	—	—	—
W8A8 [76] (QAT)	58.48	—	—	—/89.56	90.62	—/87.96	68.78	92.24	89.04/—	—	—
W8A8 (QAT)	61.21	84.80	84.64	83.82/88.85	91.29	91.29/88.28	71.12	92.89	88.39/88.18	83.37	2900
W8A8 (PTQ)	56.06	79.99	81.06	75.49/79.67	87.35	89.92/86.82	48.38	91.40	86.58/86.44	77.41	6
W8A8/16 [6] (PTQ)*	58.63	82.67	82.67	88.74	90.41	89.40	68.95	92.66	88.00	82.46	Unknown
W8A8 (ZeroQuant)	59.59	84.83	85.13	86.03/90.39	91.98	91.45/88.46	71.12	93.12	90.09/89.62	83.75	0
W4/8A16 (PTQ)	0.00	16.74	16.95	31.62/0.00	50.74	63.18/0.00	47.29	70.64	16.48/15.91	33.11	6
W4/8A16 (ZeroQuant)	57.29	82.69	83.27	84.56/88.40	90.04	86.52/79.49	70.76	92.78	88.46/88.61	81.65	0
W4/8A16 (ZeroQuant-LKD)	58.50	83.16	83.69	84.80/89.31	90.83	88.94/84.12	70.04	92.78	88.49/88.67	82.35	31
W4/8A8 (ZeroQuant)	56.69	82.46	83.06	84.07/88.03	90.13	87.04/80.50	70.76	92.78	88.07/88.44	81.55	0
W4/8A8 (ZeroQuant-LKD)	58.80	83.09	83.65	85.78/89.90	90.76	89.16/84.85	71.84	93.00	88.16/88.55	82.71	31

Mixed-precision results with LKD

- Without tuning, LKD
  - ~1 point gain with 31s on BERT-base
  - >3% acc and 50 PPL gain on GPT-3-350M

Zero-shot Eval Performance of GPT-3-350M

Precision (Method)	Lambada (↑)	PIQA (↑)	OpenBookQA (↑)	RTE (↑)	ReCoRd (↑)	Ave. 19 Tasks (↑)	Wikitext-2 (↓)	Time Cost
W16A16	49.3	66.3	29.4	53.8	75.1	38.9	21.5	N/A
W8A8 (PTQ)	42.6	64.1	28.0	53.1	67.5	37.8	26.2	7 mins
W8A8 (ZeroQuant)	51.0	66.5	29.2	53.4	74.9	38.7	21.7	0
W4/8A16 (PTQ)	0.00	51.4	30.2	52.7	16.1	28.9	1.76e5	7 mins
W4/8A16 (ZeroQuant)	10.1	58.5	27.2	52.0	56.5	33.5	88.6	0
W4/8A16 (ZeroQuant-LKD)	39.8	63.8	29.4	53.1	70.1	37.0	30.6	1.1 hours
W4/8A8 (ZeroQuant)	10.5	57.7	28.0	52.7	55.3	33.4	92.1	0
W4/8A8 (ZeroQuant-LKD)	37.4	61.8	28.2	53.1	68.5	36.6	31.1	1.1 hours

Mixed-precision results with LKD

- Without tuning, LKD
  - ~1 point gain with 31s on BERT-base
  - >3% acc and 50 PPL gain on GPT-3-350M
- With tuning (LR and Iter)
  - Extra ~0.5 gain with in total 36 GPU hours for all tasks on BERT-Base

Precision (Method)	CoLA	MNLI-m	MNLI-mm	MRPC	QNLI	QQP	RTE	SST-2	STS-B	Ave.
W16A16 (Baseline)	59.72	84.94	85.06	86.27/90.57	92.15	91.51/88.56	72.20	93.23	90.06/89.59	83.95
W8A8 (ZeroQuant-LKD No Tuning)	59.59	84.83	85.13	86.03/90.39	91.98	91.45/88.46	71.12	93.12	90.09/89.62	83.75
W8A8 (ZeroQuant-LKD Tuned)	60.90	84.95	85.10	86.27/90.60	92.07	91.47/88.47	71.84	93.46	90.09/89.62	84.07
W4/8A16 (ZeroQuant-LKD No Tuning)	58.50	83.16	83.69	84.80/89.31	90.83	88.94/84.12	70.04	92.78	88.49/88.67	82.35
W4/8A16 (ZeroQuant-LKD Tuned)	60.04	83.64	84.31	85.78/89.53	91.01	90.66/87.26	71.84	93.12	88.68/88.79	83.26
W4/8A8 (ZeroQuant-LKD No Tuning)	58.80	83.09	83.65	85.78/89.90	90.76	89.32/84.85	71.84	93.00	88.16/88.55	82.71
W4/8A8 (ZeroQuant-LKD Tuned)	60.30	83.47	84.03	85.78/89.90	90.87	90.77/87.38	71.84	93.00	88.38/88.70	83.22

LKD without original training data

- A good data resource can provide similar model accuracy
- Random data gives accuracy boost compared to ZeroQuant (no LKD)

Zero-shot Eval Performance of GPT-3-350M with W4/8A8

Method	Data Resource	Lambada (↑)	PIQA (↑)	OpenBookQA (↑)	RTE (↑)	ReCoRd (↑)	Ave. 19 Tasks (↑)	Wikitext-2 (↓)
ZeroQuant	—	10.5	57.7	28.0	52.7	55.3	33.4	92.1
ZeroQuant-LKD	Random data	26.1	59.3	29.2	50.5	64.9	34.5	40.6
ZeroQuant-LKD	Wikipedia	33.9	62.4	28.0	52.7	69.5	36.2	30.4
ZeroQuant-LKD	Original data	37.4	61.8	28.2	53.1	68.5	36.6	31.1

Outline

- ZeroQuant: Efficient and Affordable Post-Training Quantization for Large-Scale Transformers
- ZeroQuant-V2: Exploring Post-training Quantization in LLMs from Comprehensive Study to Low Rank Compensation
- ZeroQuant-FP: A Leap Forward in LLMs Post-Training W4A8 Quantization Using Floating-Point Formats

[1] [2206.01861] ZeroQuant: Efficient and Affordable Post-Training Quantization for Large-Scale Transformers (arxiv.org)
[2] https://arxiv.org/pdf/2303.08302.pdf
[3] https://arxiv.org/pdf/2307.09782.pdf

Try to Understand two things …

- What method is better for LLM PTQ?
  - Particularly, GPTQ and LKD
- Can existing methods push LLMs to even lower precision?
  - E.g., W3A16 or W2A16

Comparison between GPTQ and LKD

- GPTQ solves $\min_{\hat{W}} \|Wx - \hat{W}x\|_2^2$ using second order methods
- ZeroQuant-Global solves $\min_{\hat{\theta}} \|f_\theta(x) - f_{\hat{\theta}}(x)\|_2^2$ using LKD for an layer
- ZeroQuant-Local solves $\min_{\hat{W}} \|Wx - \hat{W}x\|_2^2$ using LKD

Precision	Method	OPT-6.7b	OPT-13b	OPT-30b	OPT-66b	BLM-1.7b	BLM-3b	BLM-7.1b	BLM-176b
W16A16		11.90	11.22	10.70	10.33	20.43	17.58	14.96	10.90
W4A16	RTN	13.44	12.09	11.52	31.52	22.47	19.01	15.90	11.20
	GPTQ	12.28	11.42	10.78	10.52	21.58	18.33	15.50	11.02
	ZQ-Local*	12.46	11.64	11.05	10.79	21.70	18.50	15.55	11.11
	ZQ-Global*	12.38	11.62	11.04	10.68	21.38	18.33	15.52	11.05
W4A8	RTN	14.80	26.36	86.26	815.00	22.75	19.17	16.19	12.22
	GPTQ	13.88	17.28	20.71	648.69	21.71	18.44	15.75	11.86
	ZQ-Local*	13.24	14.23	18.53	16.32	21.86	18.66	15.75	11.19
	ZQ-Global*	13.17	13.07	14.65	37.82	21.43	18.39	15.58	11.49

Comparison between GPTQ and LKD

- GPTQ works better for weight-only quantization
- ZeroQuant works better for weight&activation quantization

Precision	Method	OPT-6.7b	OPT-13b	OPT-30b	OPT-66b	BLM-1.7b	BLM-3b	BLM-7.1b	BLM-176b
W16A16		11.90	11.22	10.70	10.33	20.43	17.58	14.96	10.90
W4A16	RTN	13.44	12.09	11.52	31.52	22.47	19.01	15.90	11.20
	GPTQ	12.28	11.42	10.78	10.52	21.58	18.33	15.50	11.02
	ZQ-Local*	12.46	11.64	11.05	10.79	21.70	18.50	15.55	11.11
	ZQ-Global*	12.38	11.62	11.04	10.68	21.38	18.33	15.52	11.05
W4A8	RTN	14.80	26.36	86.26	815.00	22.75	19.17	16.19	12.22
	GPTQ	13.88	17.28	20.71	648.69	21.71	18.44	15.75	11.86
	ZQ-Local*	13.24	14.23	18.53	16.32	21.86	18.66	15.75	11.19
	ZQ-Global*	13.17	13.07	14.65	37.82	21.43	18.39	15.58	11.49

Extra-low precision

- Either GPTQ or LKD can make extra-low precision work

Bits	Coarse-grained weight quantization (per-row block-size)				
	OPT-6.7b	OPT-13b	OPT-30b	OPT-66b	BLM-176b
W8A16	11.90	11.22	10.70	10.33	10.90
W4A16	12.28	11.42	10.78	10.78	11.02
W3A16	14.18	12.43	11.28	17.77	49.46
W2A16	120.56	40.17	25.74	225.45	Explode

Extra-low precision

- Either GPTQ or LKD can make extra-low precision work
- Low rank compensation (LoRC)
  - $W_{final} = W_{quant} + UV$, where $UV = SVD(W\text{-}W_{quant})$ with rank=8

Bits	Coarse-grained weight quantization (per-row block-size)				
	OPT-6.7b	OPT-13b	OPT-30b	OPT-66b	BLM-176b
W8A16	11.90	11.22	10.70	10.33	10.90
W4A16	12.28	11.42	10.78	10.78	11.02
W3A16	14.18	12.43	11.28	17.77	49.46
W2A16	120.56	40.17	25.74	225.45	Explode

Extra-low precision

- Either GPTQ or LKD can make extra-low precision work
- Low rank compensation (LoRC)
  - $W_{final} = W_{quant} + UV$, where $UV = SVD(W\text{-}W_{quant})$ with rank=8

Bits	LoRC	Coarse-grained weight quantization (per-row block-size) OPT-6.7b	OPT-13b	OPT-30b	OPT-66b	BLM-176b
W8A16		11.90	11.22	10.70	10.33	10.90
W4A16	✗	12.28	11.42	10.78	10.78	11.02
	✓	12.10	11.36	10.76	10.34	10.98
W3A16	✗	14.18	12.43	11.28	17.77	49.46
	✓	13.00	11.90	11.14	10.63	11.30
W2A16	✗	120.56	40.17	25.74	225.45	Explode
	✓	24.17	18.53	14.39	13.01	14.15

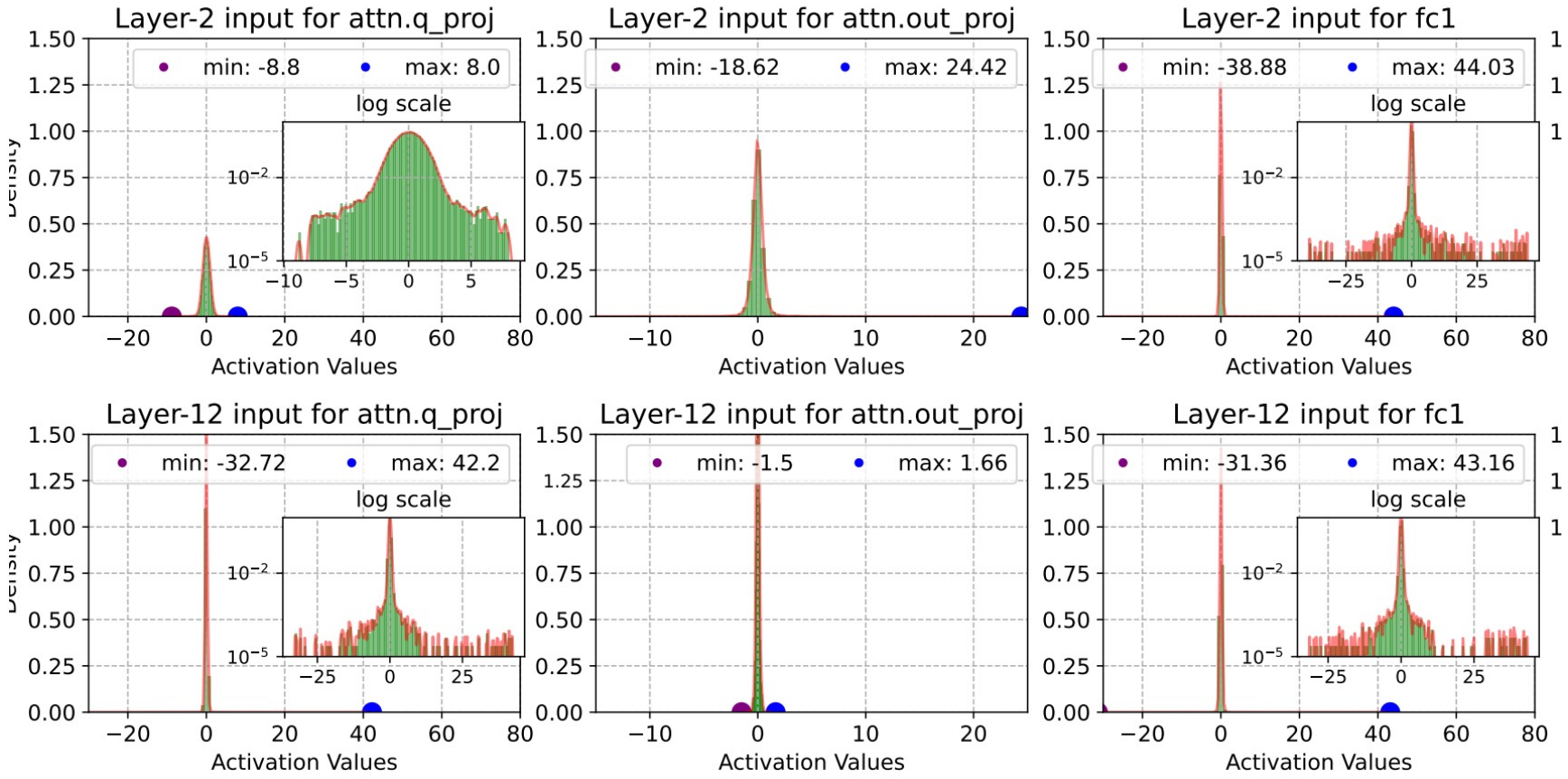
Outline

- ZeroQuant: Efficient and Affordable Post-Training Quantization for Large-Scale Transformers
- ZeroQuant-V2: Exploring Post-training Quantization in LLMs from Comprehensive Study to Low Rank Compensation
- ZeroQuant-FP: A Leap Forward in LLMs Post-Training W4A8 Quantization Using Floating-Point Formats

[1] [2206.01861] ZeroQuant: Efficient and Affordable Post-Training Quantization for Large-Scale Transformers (arxiv.org)
[2] https://arxiv.org/pdf/2303.08302.pdf
[3] https://arxiv.org/pdf/2307.09782.pdf

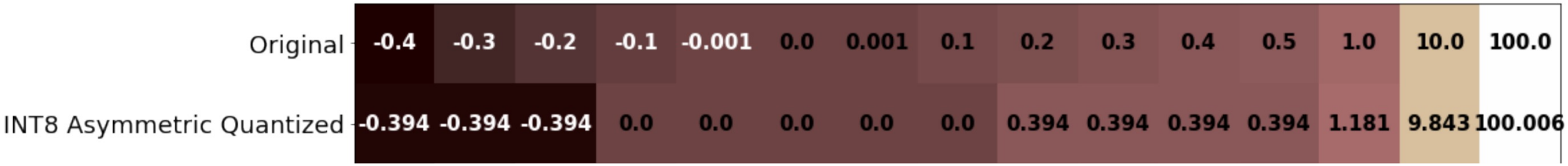
How can we quantize activation?

- Activation's outlier is the killer for quantization



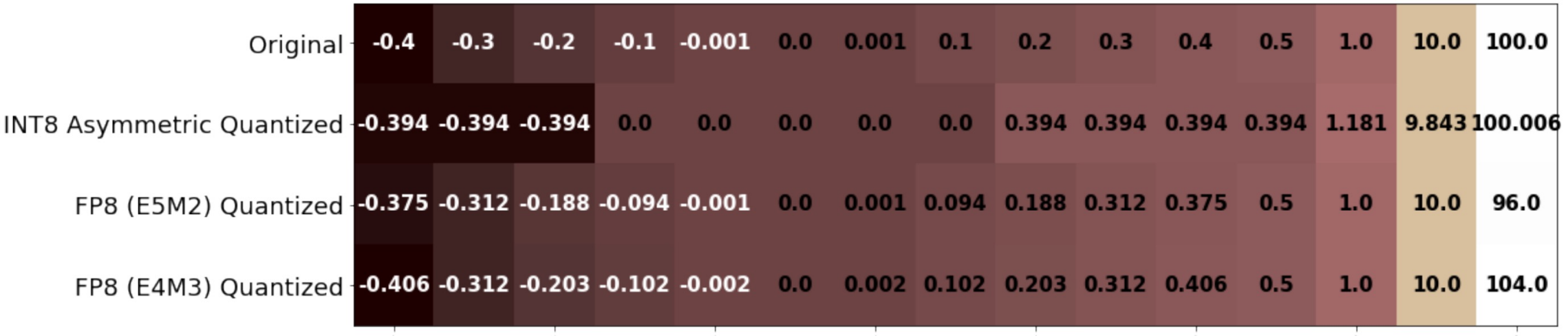
How can we quantize activation?

- Activation’s outlier is the killer for quantization
  - INT format: capture outliers but lose accurate representation for small values



How can we quantize activation?

- Activation's outlier is the killer for quantization
  - INT format: capture outliers but lose accurate representation for small values
- Floating point format is designed for dense+outlier distribution



FP vs. INT results

- OPT families
  - INT+INT does not work; X+INT works

Q-type	Weight – – Activation	Mean	OPT-3b WIKI/PTB/C4	Mean	OPT-7b WIKI/PTB/C4	Mean	OPT-13b WIKI/PTB/C4	Mean	OPT-30b WIKI/PTB/C4
W16A16	N/A	15.44	14.62/16.97/14.72	11.90	10.86/13.09/11.74	11.22	10.13/12.34/11.20	10.70	9.56/11.84/10.69
W8A8	INT – INT	15.94	14.98/17.49/15.36	12.66	11.20/14.29/12.48	15.94	12.13/19.82/15.86	25.76	14.63/32.90/29.74
	INT – FP	15.85	14.93/17.56/15.05	11.99	10.92/13.24/11.80	11.27	10.16/12.42/11.23	10.69	9.51/11.87/10.71
	FP – FP	15.86	14.97/17.55/15.05	11.99	10.91/13.24/11.81	11.27	10.16/12.42/11.23	10.69	9.51/11.87/10.71
W4A8	INT – INT	16.41	15.39/18.22/15.62	13.18	11.61/15.00/12.92	16.70	12.32/21.21/16.56	24.42	14.80/30.38/28.09
	INT – FP	16.40	15.46/18.23/15.51	12.20	11.13/13.49/11.99	11.34	10.20/12.53/11.30	10.73	9.54/11.91/10.75
	FP – FP	16.29	15.32/18.19/15.35	12.09	10.89/13.44/11.95	11.34	10.16/12.55/11.30	10.72	9.52/11.90/10.75

FP vs. INT results

- OPT families
  - INT+INT does not work; X+INT works
- LLaMa families
  - INT+INT < X+FP

Q-type	Weight-Activation	LLaMA-3b Mean	LLaMA-3b WIKI/PTB/C4	LLaMA-7b Mean	LLaMA-7b WIKI/PTB/C4	LLaMA-13b Mean	LLaMA-13b WIKI/PTB/C4	LLaMA-30b Mean	LLaMA-30b WIKI/PTB/C4
W16A16	N/A	11.93	7.35/19.1/9.34	13.37	5.68/27.35/7.78	10.31	5.09/19.22/6.61	5.79	4.10/7.30/5.98
W8A8	INT – INT	12.00	7.41/19.16/9.41	13.58	5.72/27.89/7.13	10.63	5.16/20.07/6.67	5.90	4.21/7.42/6.06
	INT – FP	11.96	7.37/19.16/9.35	13.45	5.69/27.57/7.09	10.38	5.11/19.42/6.62	5.80	4.11/7.31/5.99
	FP – FP	11.99	7.37/19.23/9.37	13.46	5.70/27.58/7.10	10.38	5.11/19.41/6.62	5.81	4.12/7.31/5.99
W4A8	INT – INT	12.55	7.67/20.23/9.74	16.23	6.44/34.45/7.79	11.48	5.32/22.35/6.78	6.02	4.36/7.54/6.16
	INT – FP	12.39	7.62/19.87/9.68	16.09	6.75/33.80/7.72	11.31	5.28/21.91/6.73	5.94	4.27/7.45/6.11
	FP – FP	12.45	7.62/20.05/9.67	15.14	6.32/31.61/7.51	11.08	5.26/21.27/6.73	5.92	4.26/7.42/6.09

Thank You for Listening!

zheweiyao@gmail.com