CS 498: Machine Learning System Spring 2025

Minjia Zhang

The Grainger College of Engineering

Today

DL Inference

- LLM Quantization

Objective: Understand how LLM quantization reduces memory and compute cost, and learn two key methods, ZeroQuant and SmoothQuant, for compressing weights and activations.

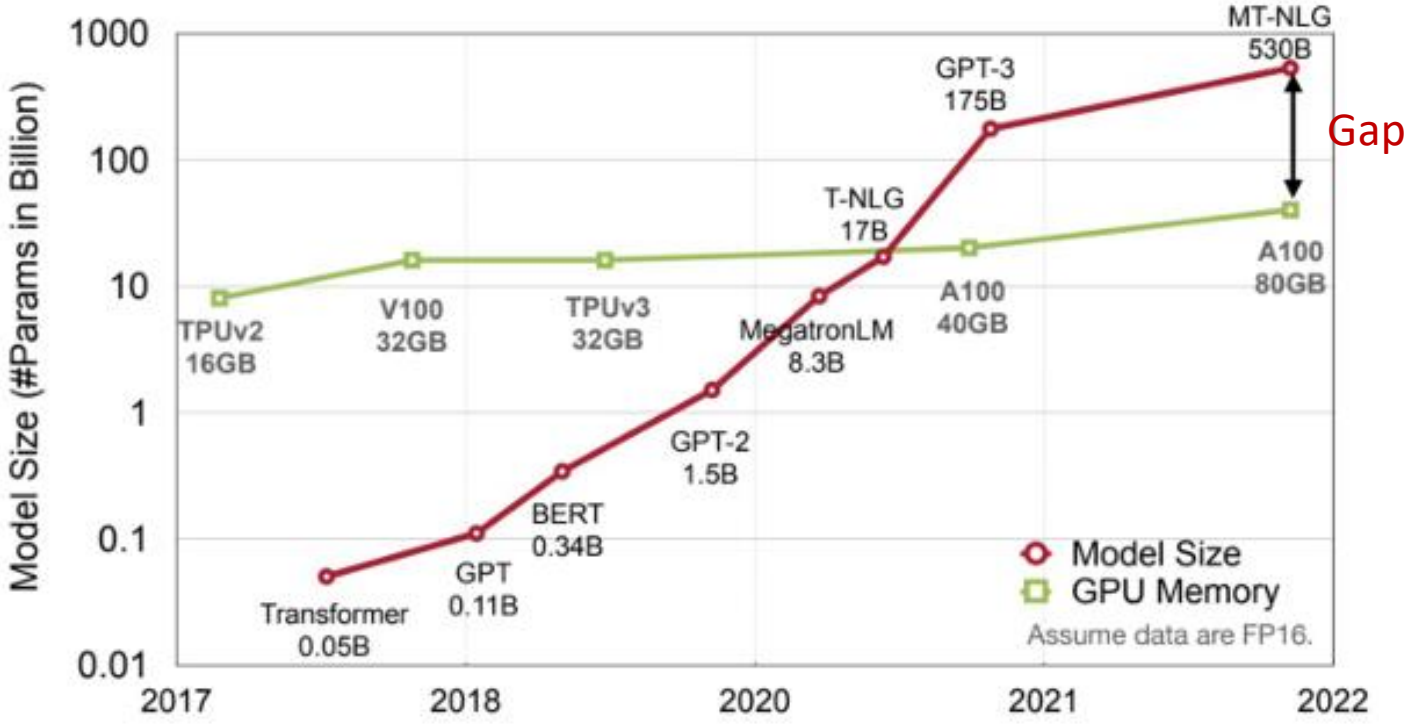
Serving Large Language Models Is Expensive

- Large language models (LLMs) are taking over every field.

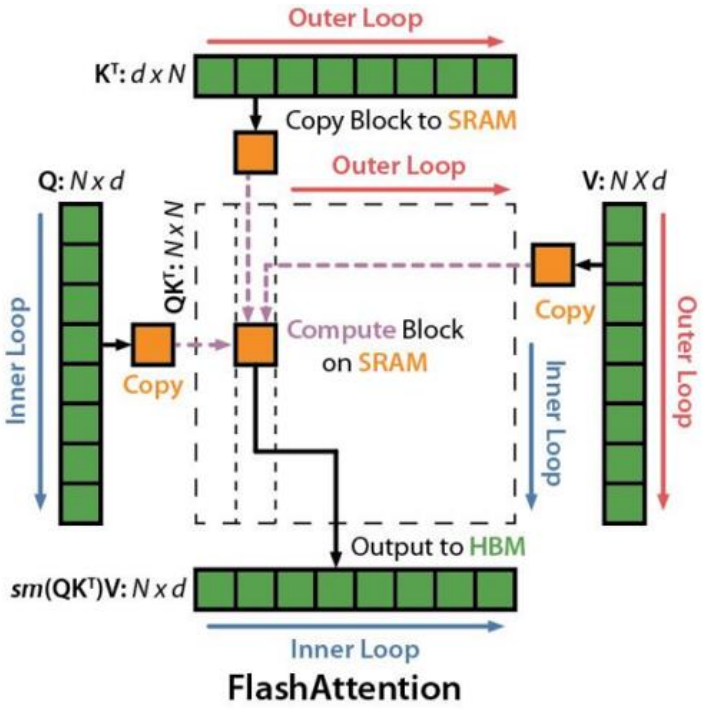
- As the models get larger, serving such models for inference becomes **expensive** and **challenging**!

Gap Between the Supply and Demand of LLMs

- Size of LLMs is increasing faster than GPU memory: **exponential growth in model size** vs. linear increase in GPU memory
- **Memory challenges:** 175B parameters in GPT-3 requires 350GB of memory to store just weights
- **Deployment difficulties:** Requires multiple GPUs, high latency, and resource intensive setups

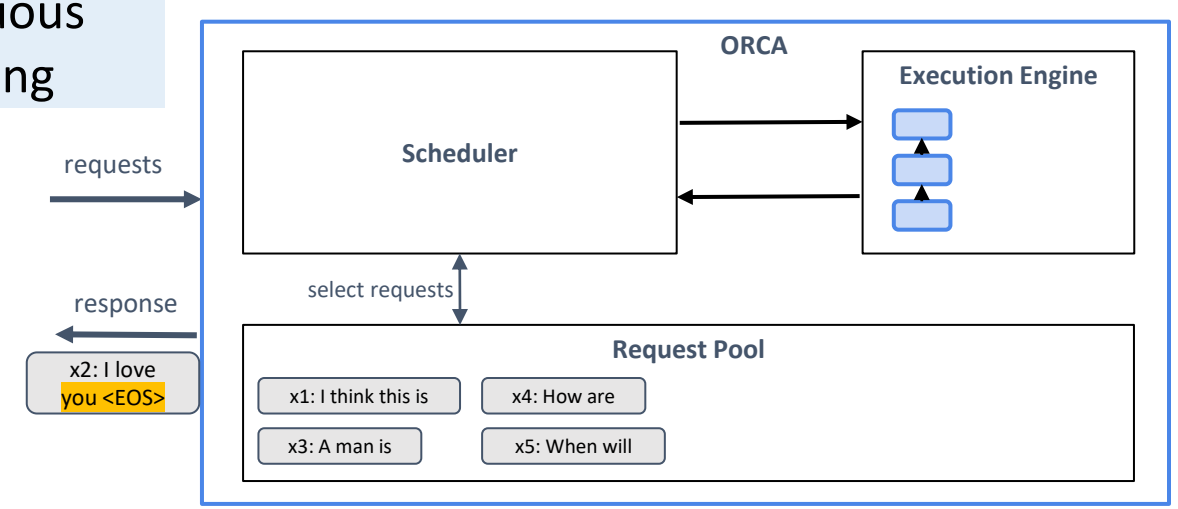


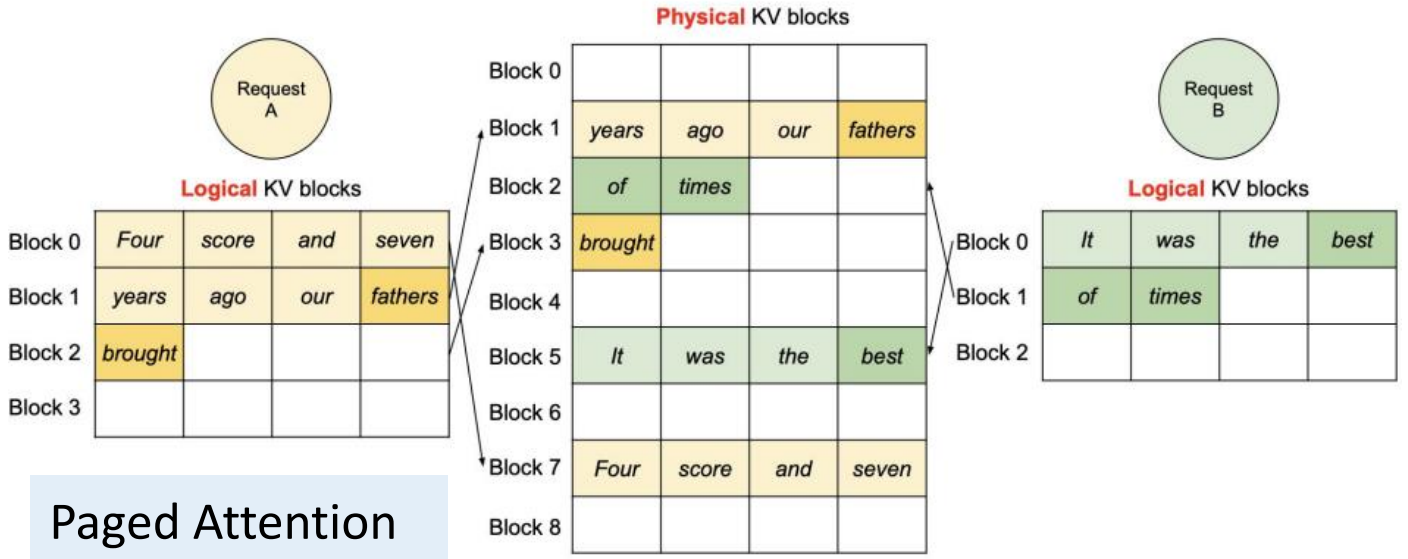
Can Existing Methods Reduce LLM Memory Consumption?



Flash Attention

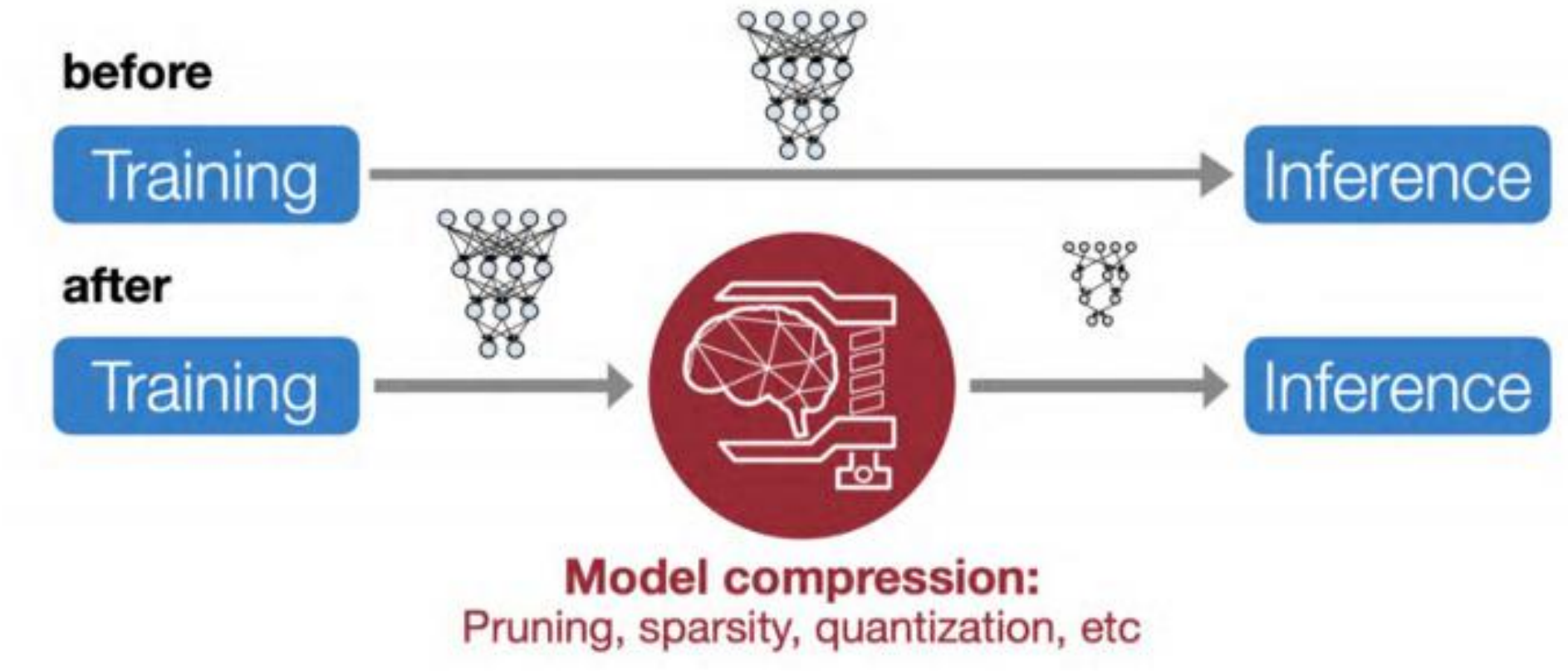
Continuous Batching





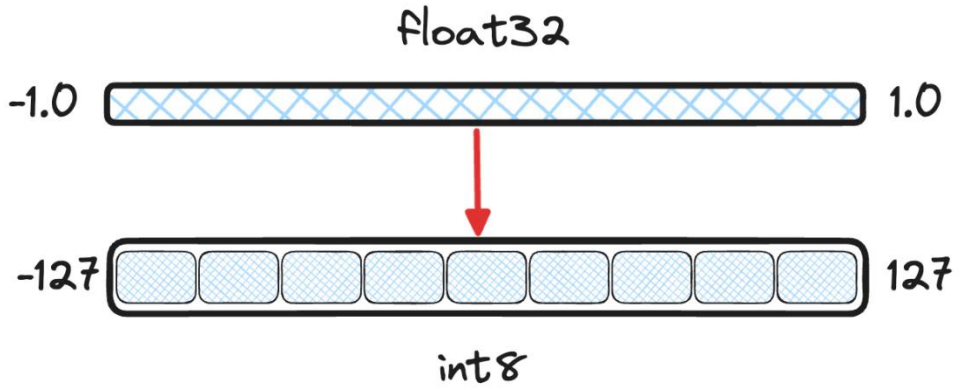
Paged Attention

Bridge the Gap through Model Compression



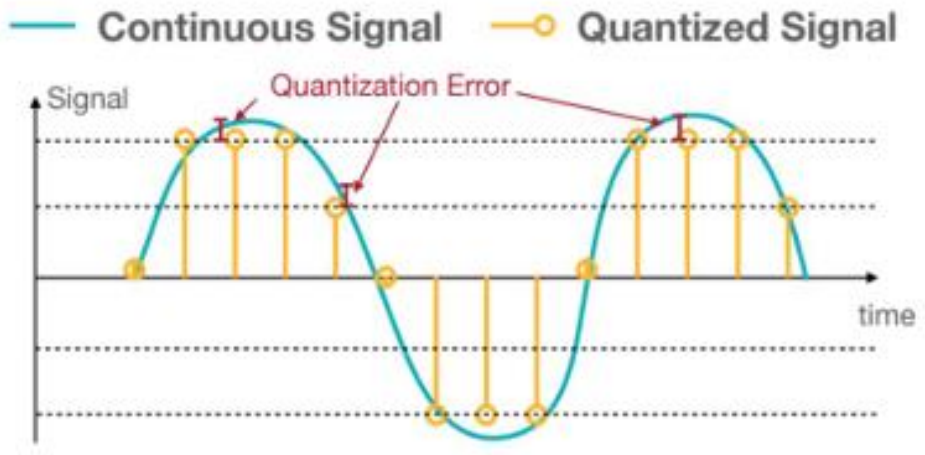
Quantization

- Reduce the bits per weight, saving memory consumption
- Accelerate inference speed on supporting hardware



What is Quantization?

- *Quantization is the process of mapping a large set of continuous or high-precision values (typically floating-point numbers) to a smaller set of discrete values (typically lower-precision integers) in order to reduce the computational and memory requirements of a model.*

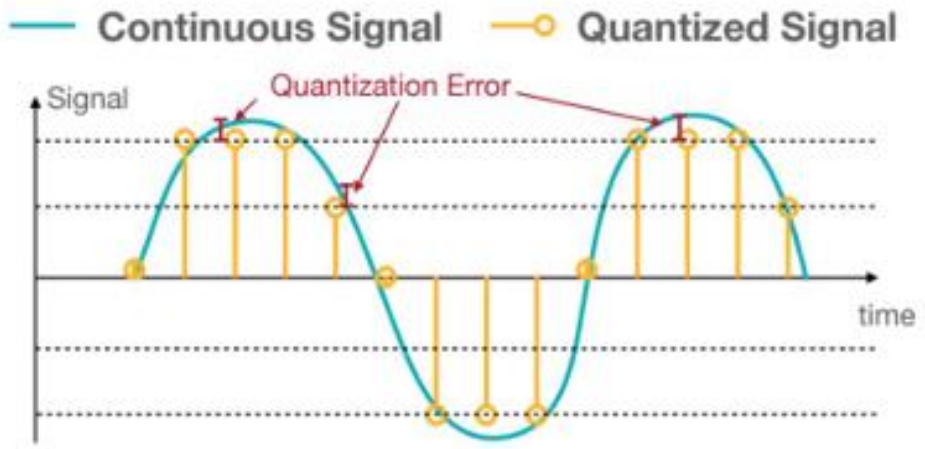


The difference between an input value and its quantized value is referred to as quantization error.

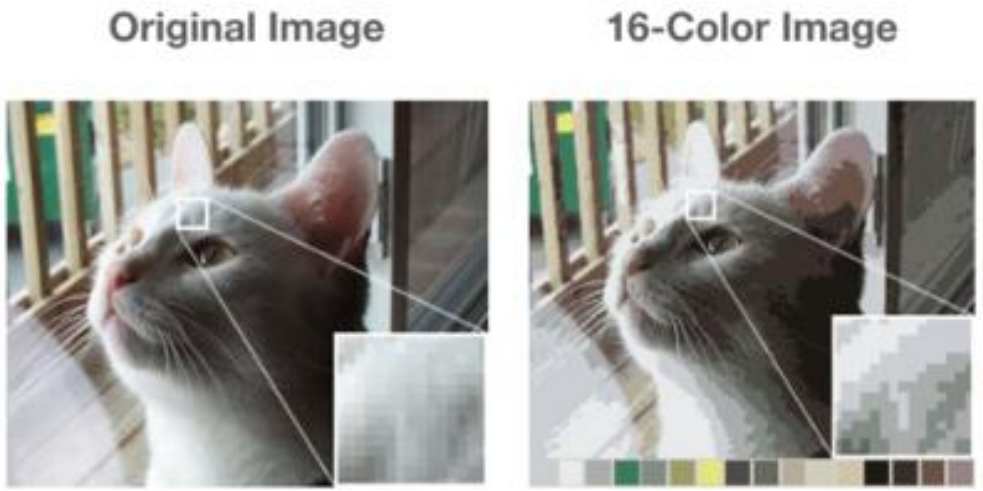
Originate from signal processing and information theory

What is Quantization?

- *Quantization is the process of mapping a large set of continuous or high-precision values (typically floating-point numbers) to a smaller set of discrete values (typically lower-precision integers) in order to reduce the computational and memory requirements of a model.*



The difference between an input value and its quantized value is referred to as quantization error.



Originate from signal processing and information theory

Adapted to computer vision models in ML

Motivation for LLM Quantization

- LLMs often come in high precision formats such as FP 16 or FP32
  - Significant GPU memory requirements (memory capacity and bandwidth)

Motivation for LLM Quantization

- LLMs often come in high precision formats such as FP 16 or BF16
  - Significant GPU memory requirements (memory capacity and bandwidth)

- Question: Do you know why pre-trained LLMs are often in FP16/BF16?

Key Concepts: Linear Quantization

- An affine mapping of integers to real numbers $r = S(q - Z)$

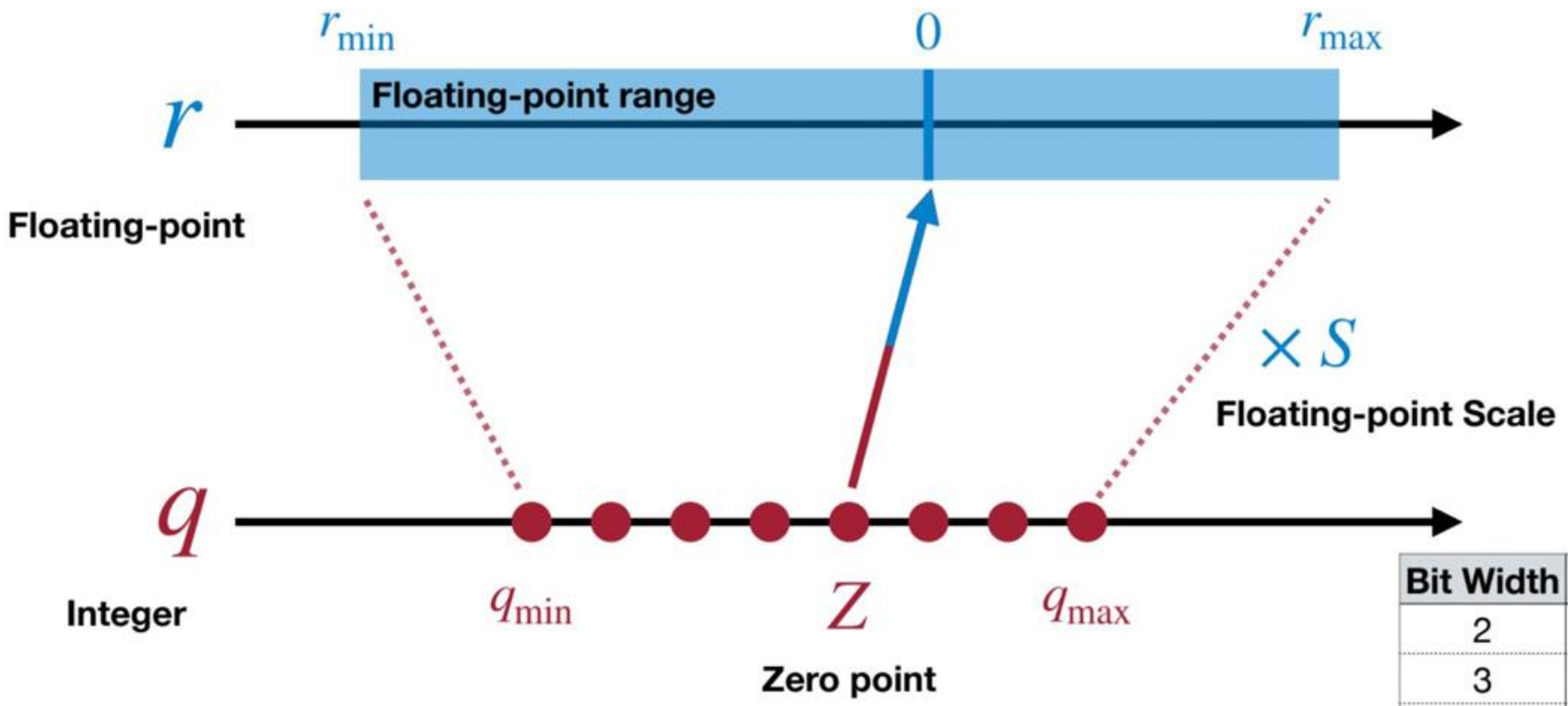
Key Concepts: Linear Quantization

- An affine mapping of integers to real numbers $r = S(q - Z)$

Bit Width	q_{min}	q_{max}
2	-2	1
3	-4	3
4	-8	7
N	-2^{N-1}	$2^{N-1}-1$

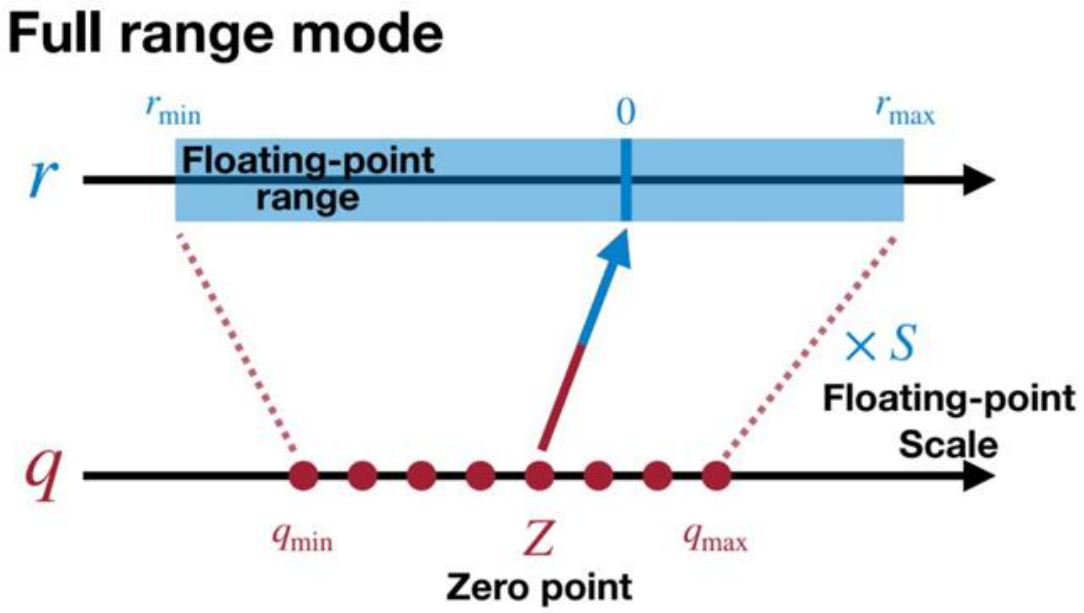
Key Concepts: Linear Quantization

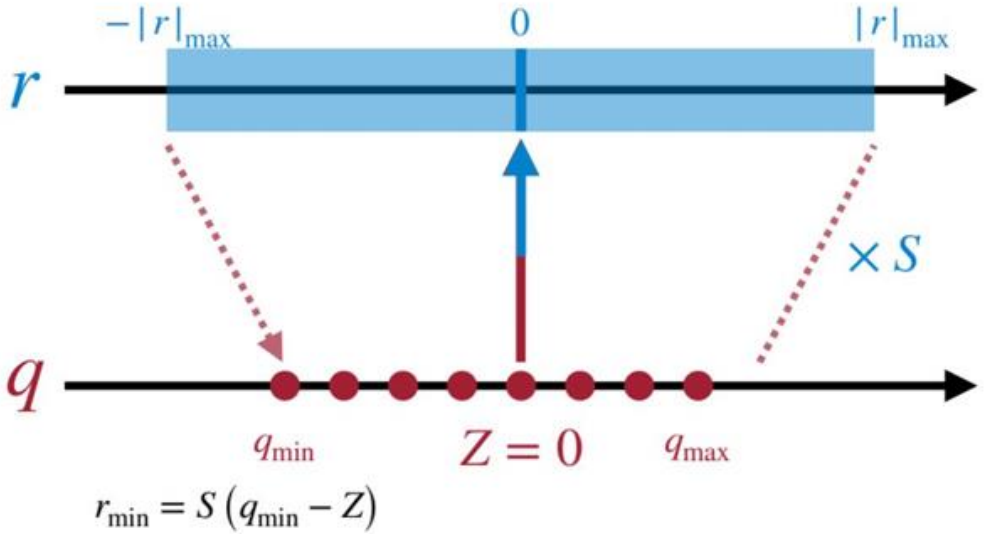
- An affine mapping of integers to real numbers $r = S(q - Z)$



Bit Width	q_{min}	q_{max}
2	-2	1
3	-4	3
4	-8	7
N	-2^{N-1}	$2^{N-1}-1$

Key Concepts: Symmetric Linear Quantization





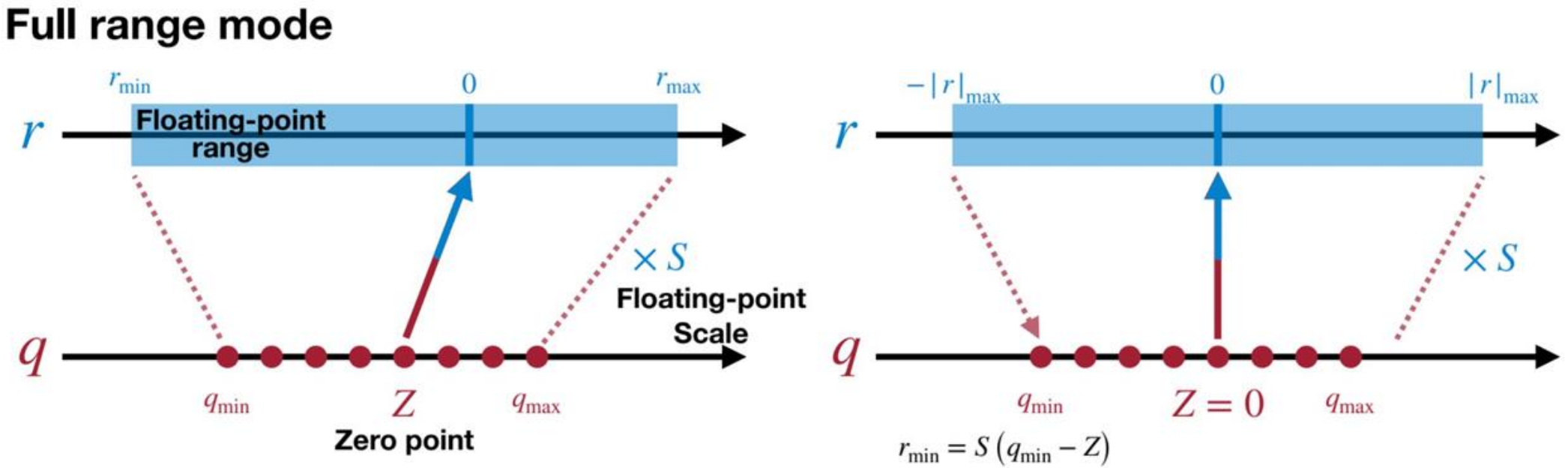
Bit Width	q_{min}	q_{max}
2	-2	1
3	-4	3
4	-8	7
N	-2^{N-1}	$2^{N-1}-1$

$$S = \frac{r_{max} - r_{min}}{q_{max} - q_{min}}$$

$$S = \frac{r_{min}}{q_{min} - Z} = \frac{-|r|_{max}}{q_{min}} = \frac{|r|_{max}}{2^{N-1}}$$

- use full range of quantized integers
- example: PyTorch's native quantization,

Key Concepts: Symmetric Linear Quantization



Bit Width	q_{min}	q_{max}
2	-2	1
3	-4	3
4	-8	7
N	-2^{N-1}	$2^{N-1}-1$

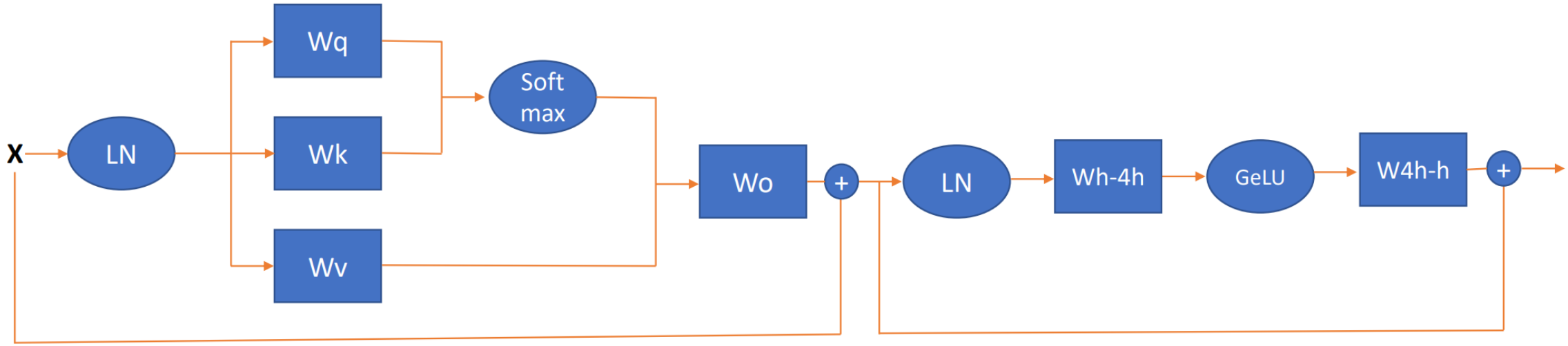
$$S = \frac{r_{max} - r_{min}}{q_{max} - q_{min}}$$

$$S = \frac{r_{min}}{q_{min} - Z} = \frac{-|r|_{max}}{q_{min}} = \frac{|r|_{max}}{2^{N-1}}$$

- use full range of quantized integers
- example: PyTorch's native quantization,

Question: How do we apply this to LLMs?

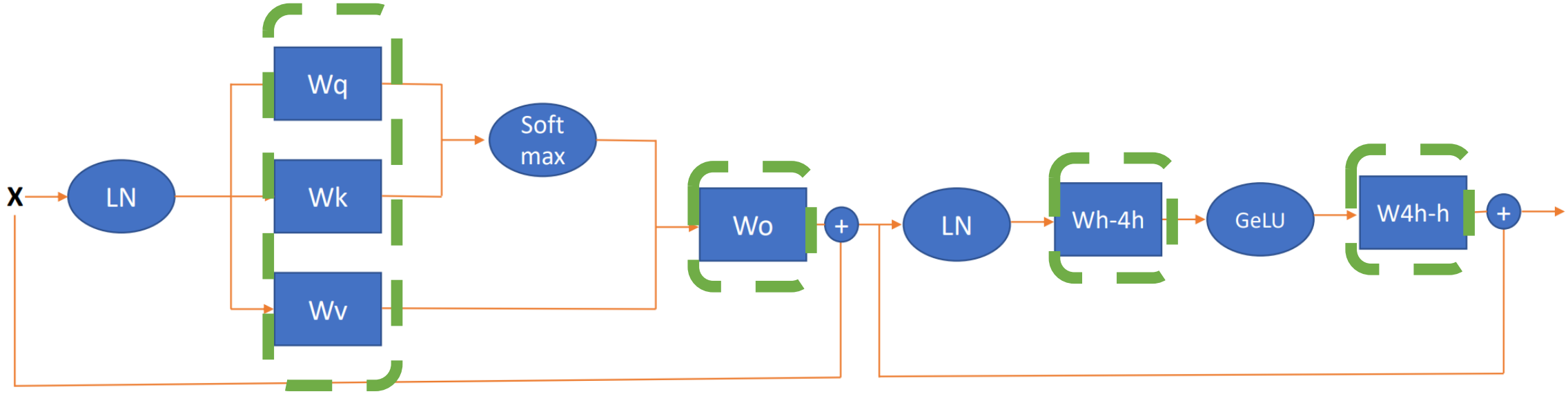
LLM 8-bit Quantization



- 8-bit quantization

$$\mathbf{x}_{quantize} = round\left(clamp(\frac{\mathbf{x}}{S}, -2^{bit-1}, 2^{bit-1}-1)\right)$$

LLM 8-bit Weight Quantization



- 8-bit weight quantization

$$\mathbf{x}_{quantize} = round\left(clamp(\frac{\mathbf{x}}{S}, -2^{bit-1}, 2^{bit-1}-1)\right)$$

FP32 weight matrix

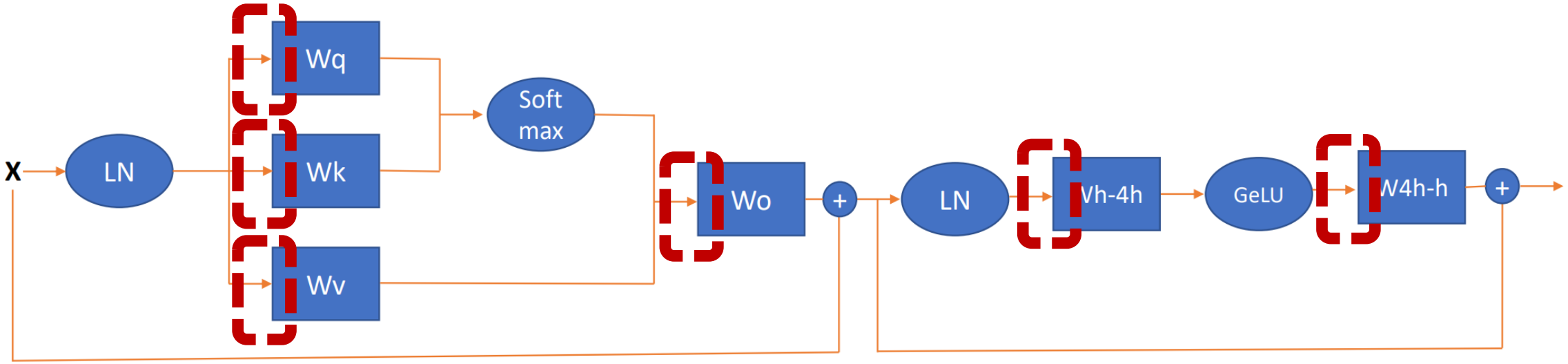
1.1	2.2	0.1	-0.1	-5.5	-6.6
…					
…					
…					
…					
1.1	2.1	0.1	-0.1	-4.8	-6.6

≈ Scaling Factor $1/S$ 0.05 *

8-bit quantization

21	42	2	-2	-106	-127
…					
…					
…					
…					
21	40	2	-2	-92	-127

LLM 8-bit Activation Quantization



- 8-bit activation
  (Input to the linear layer)

$$\mathbf{x}_{quantize} = round\left(clamp(\frac{\mathbf{x}}{S}, -2^{bit-1}, 2^{bit-1}-1)\right)$$

FP32 activation matrix

1.1	2.2	0.1	-0.1	-5.5	-6.6
…					
…					
…					
…					
1.1	2.1	0.1	-0.1	-4.8	-6.6

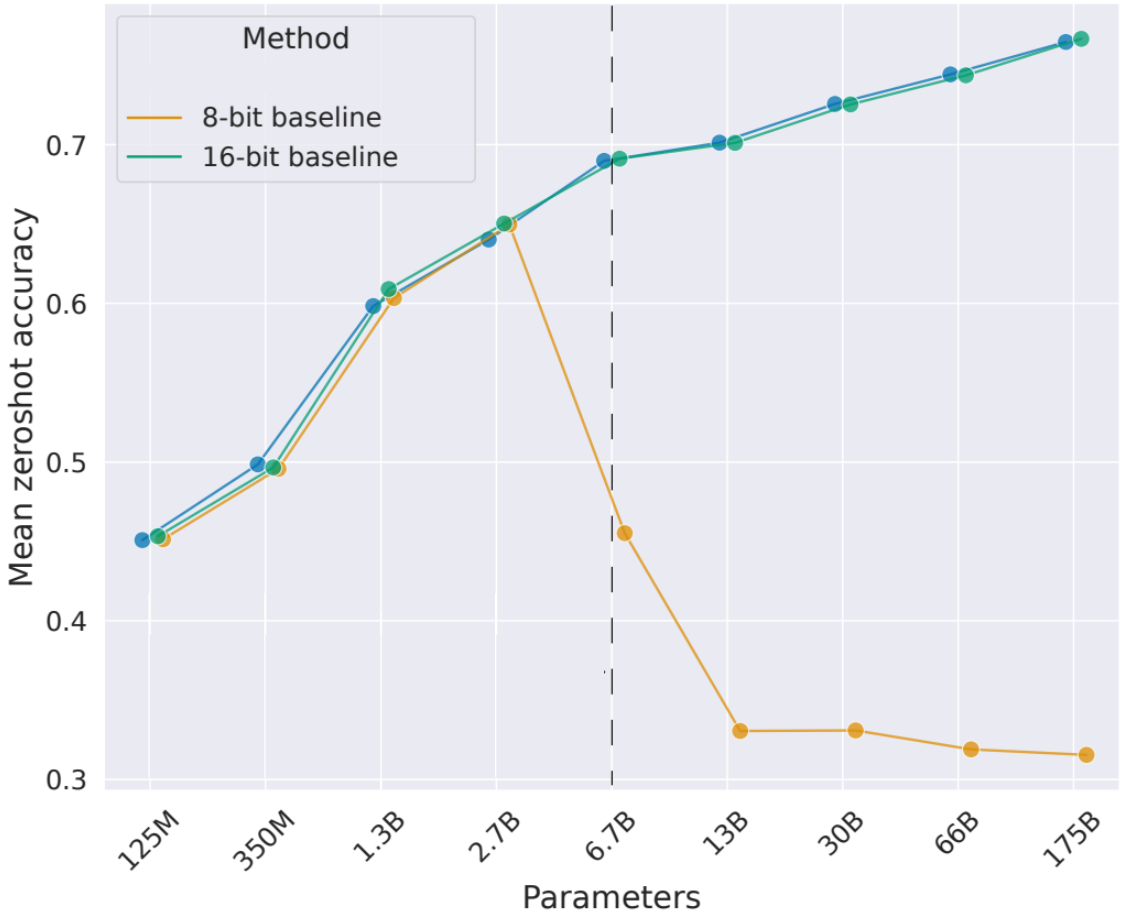
Scaling Factor $1/S$ $\approx$ 0.05*

8-bit quantization

21	42	2	-2	-106	-127
…					
…					
…					
…					
21	40	2	-2	-92	-127

LLM Quantization Challenges

- Standard quantization strategy leads to catastrophic accuracy drop

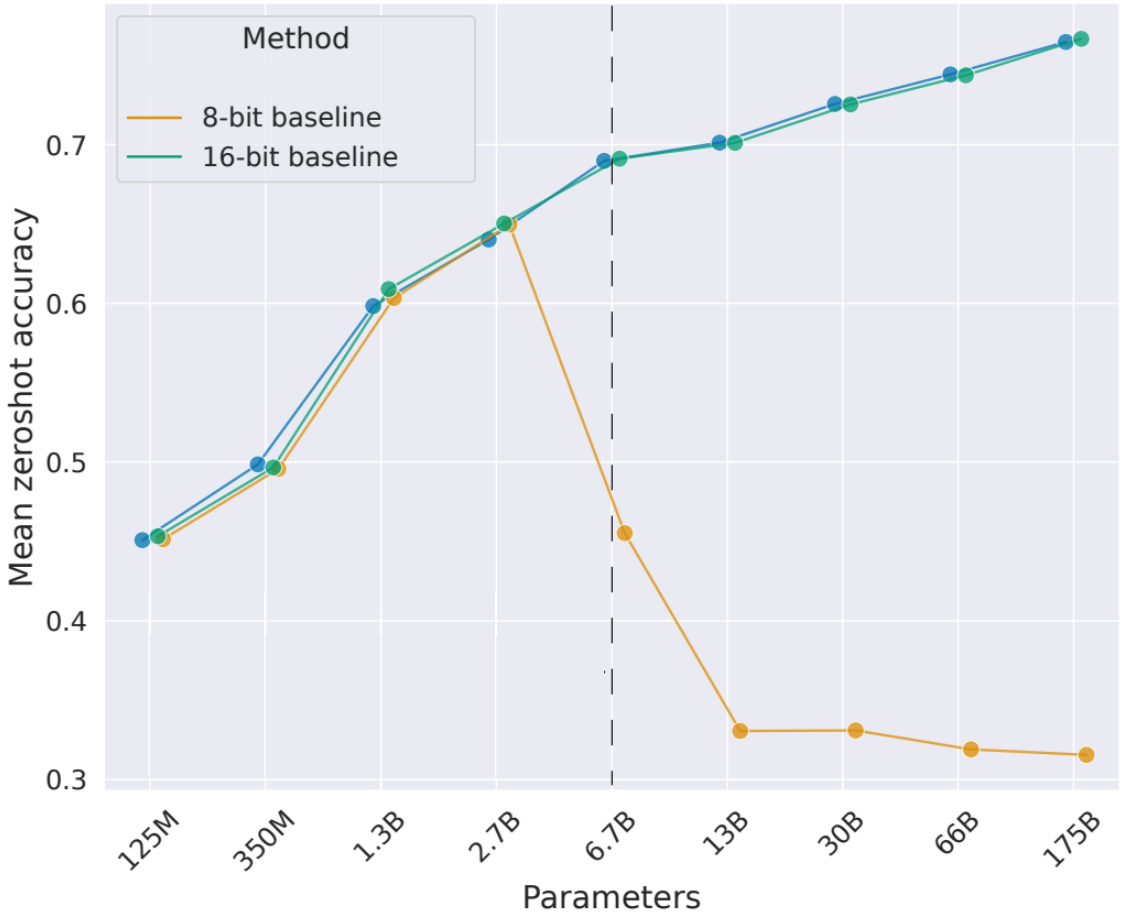


LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale, 2023

LLM Quantization Challenges

- Standard quantization strategy leads to catastrophic accuracy drop

Group Discussion: Why do you think this happens? How would you begin to diagnose this issue?



LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale, 2023

LLM Quantization Challenges

Zero-shot evaluation results. WxAy means x-/y-bit for weight/activation.

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

Question: Take a close look at the performance across different quantization settings in the table. Can you identify what causes LLM to loss accuracy?

ZeroQuant: Efficient and Affordable Post-Training Quantization for Large-Scale Transformers, 2023

LLM Quantization Challenges

Zero-shot evaluation results. WxAy means x-/y-bit for weight/activation.

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

ZeroQuant: Efficient and Affordable Post-Training Quantization for Large-Scale Transformers, 2023

LLM Quantization Challenges

Zero-shot evaluation results. WxAy means x-/y-bit for weight/activation.

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

ZeroQuant: Efficient and Affordable Post-Training Quantization for Large-Scale Transformers, 2023

LLM Quantization Challenges

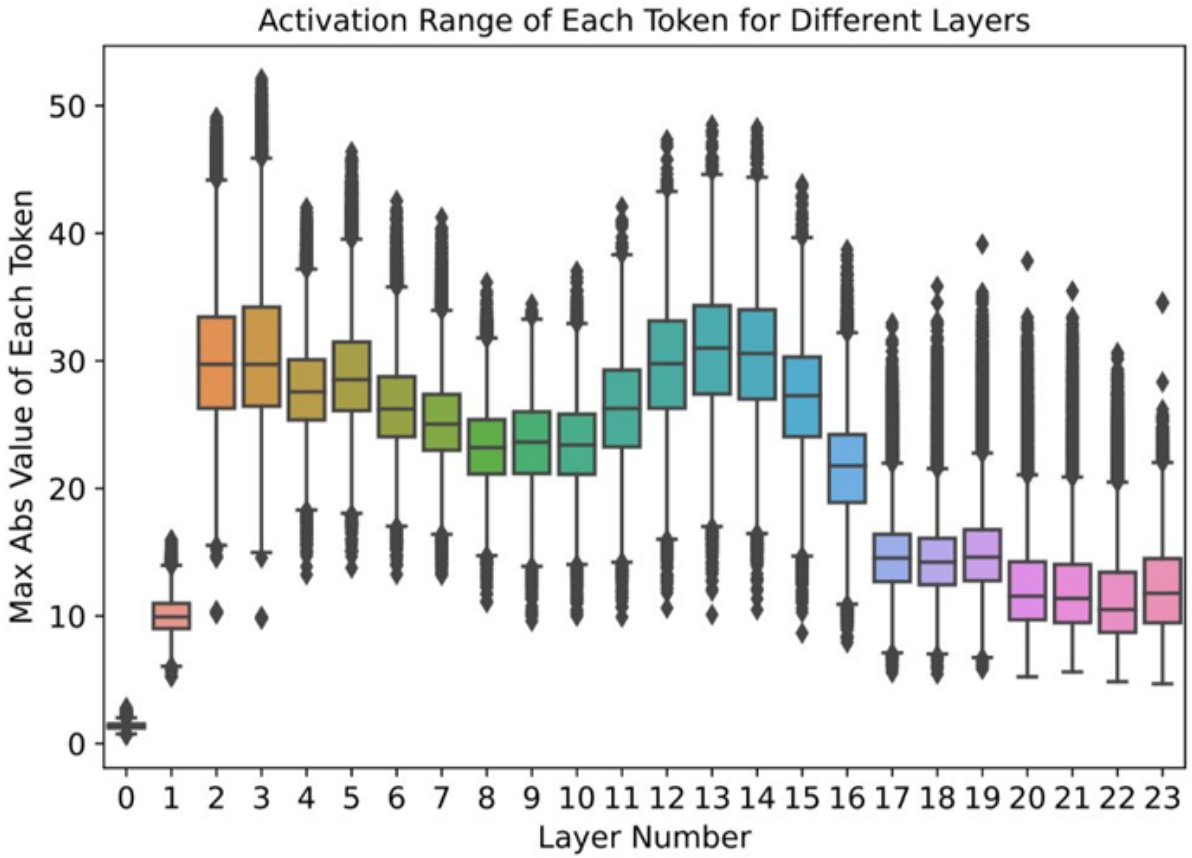
Zero-shot evaluation results. WxAy means x-/y-bit for weight/activation.

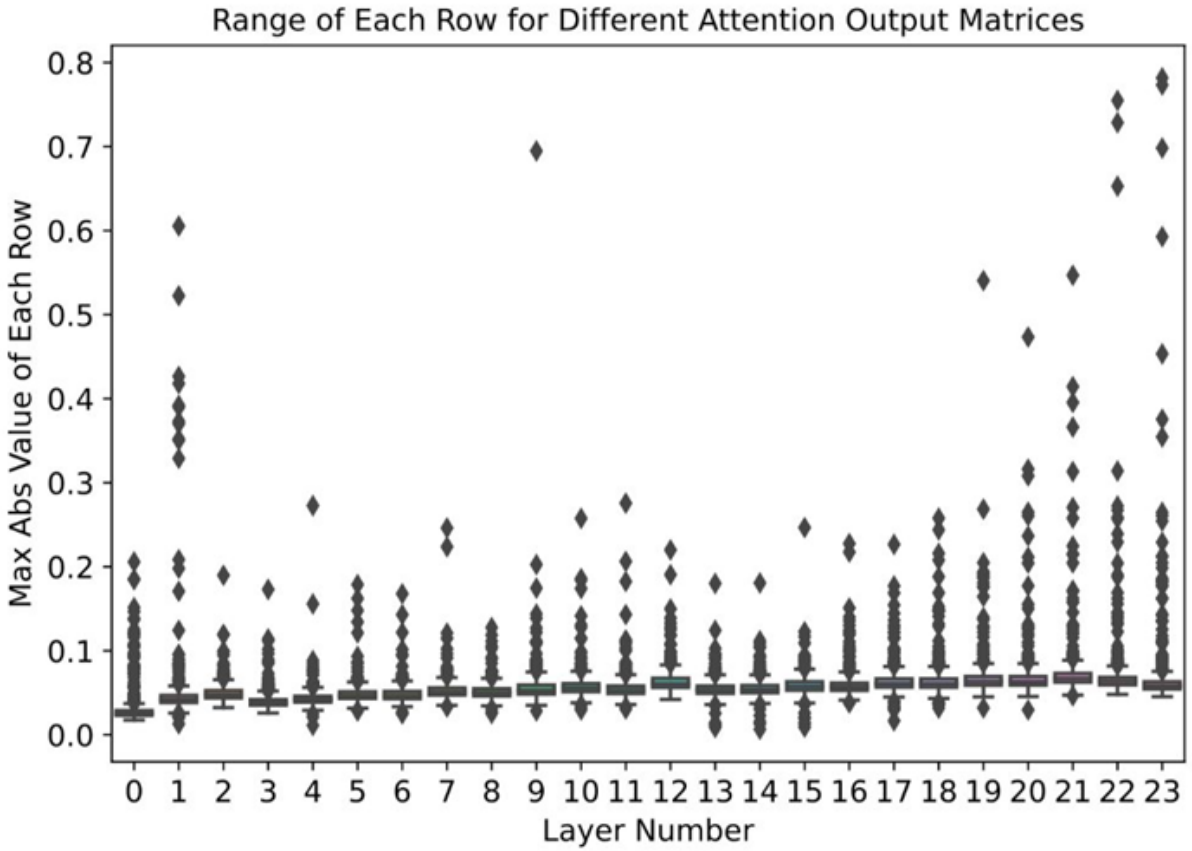
Precision	Lambada (↑)	PIQA (↑)	OpenBookQA (↑)	RTE (↑)	ReCoRd (↑)	Ave. 19 Tasks (↑)	Wikitext-2 (↓)
W16A16	49.3	66.3	29.4	53.8	75.1	38.9	21.5
W8A16	49.3	66.1	29.6	54.2	74.8	38.5	22.1
W16A8	44.7	64.8	28.2	52.7	69.2	37.8	24.6
W8A8	42.6	64.1	28.0	53.1	67.5	37.8	26.2
W4/8A16	0.00	51.4	30.2	52.7	16.1	28.9	1.76e5

INT8 activation quantization causes the primary accuracy loss

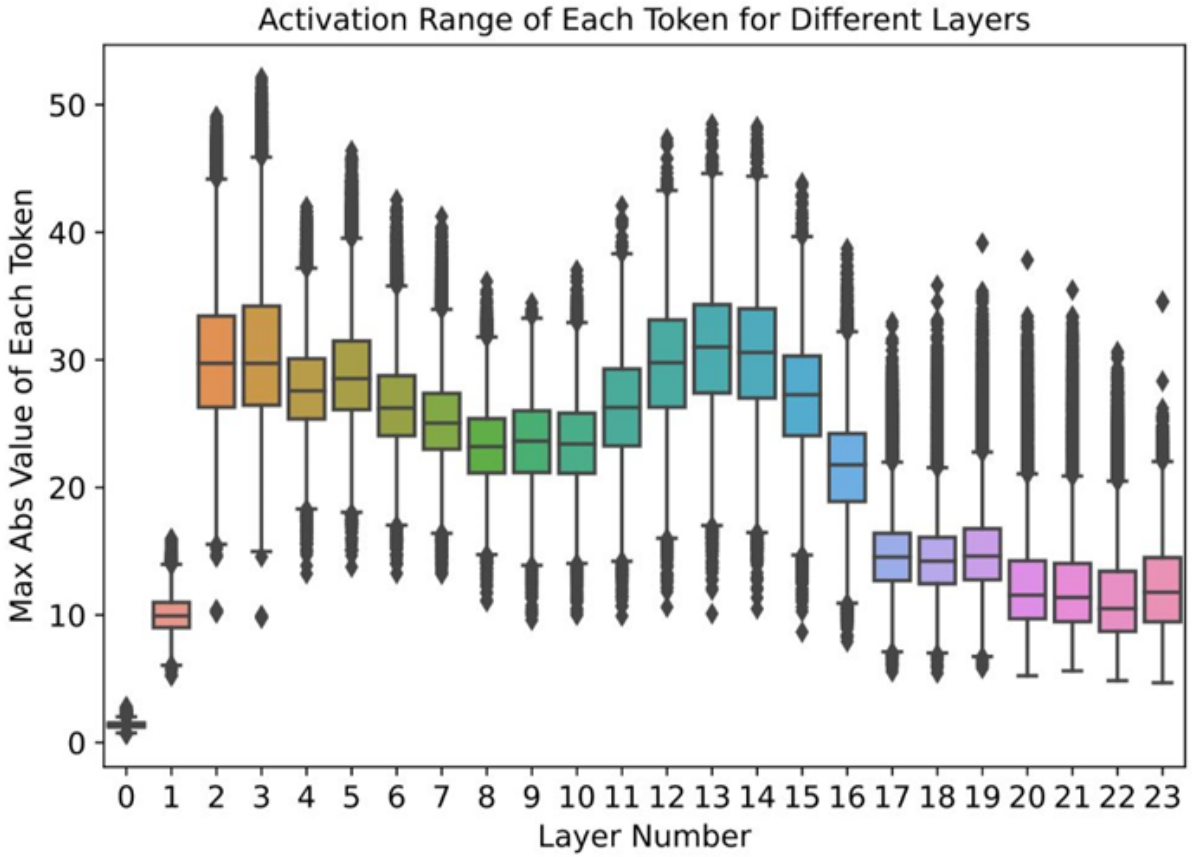
ZeroQuant: Efficient and Affordable Post-Training Quantization for Large-Scale Transformers, 2023

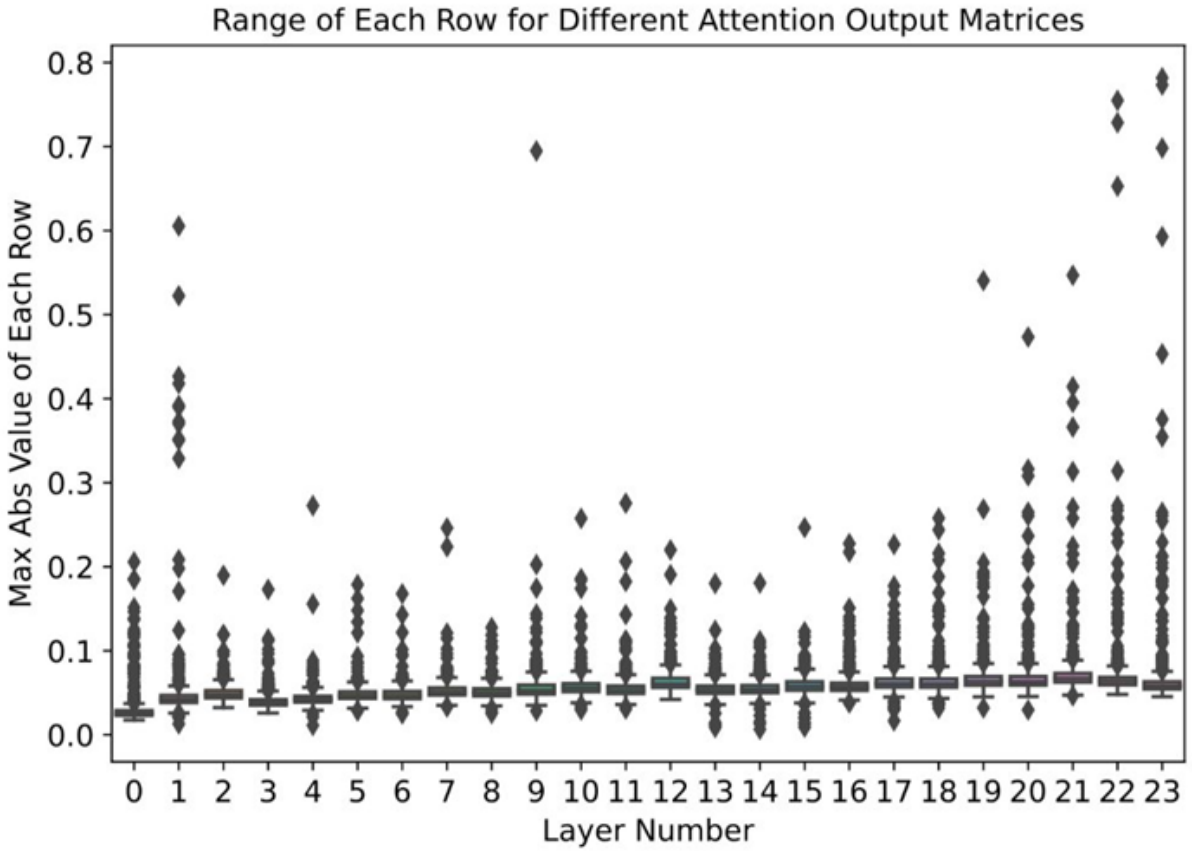
LLM Quantization Challenges





LLM Quantization Challenges



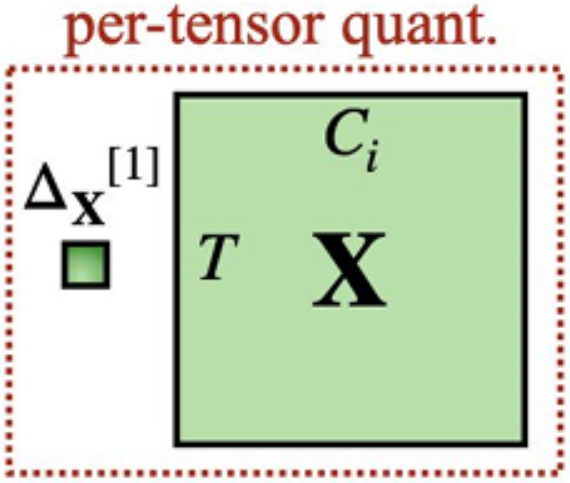


Questions: How do we mitigate accuracy loss from dynamic range (outliers)?

Quantization Granularity

- Per-Tensor Quantization
  - Uses a single step size for the entire matrix
- Per-Token Quantization
- Per-Channel Quantization
- Group-Wise Quantization

$$S = \frac{|r|_{\max}}{q_{\max}} = \frac{2.12}{2^{2-1} - 1} = 2.12$$



1	0	1	0
0	0	-1	1
0	1	0	0
1	0	1	1

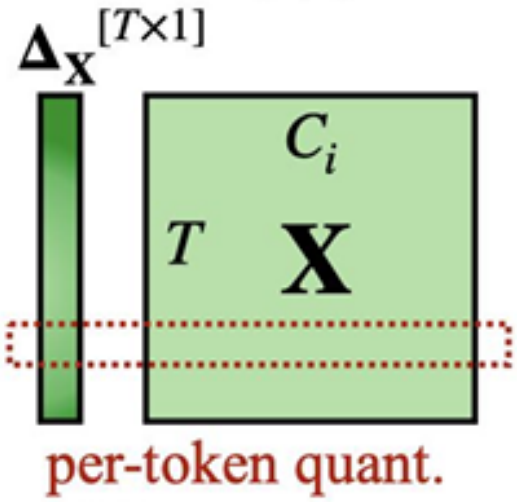
Quantized

2.12	0	2.12	0
0	0	-2.12	2.12
0	2.12	0	0
2.12	0	2.12	2.12

Reconstructed

Quantization Granularity

- Per-Tensor Quantization
- Per-Token Quantization
  - Uses different quantization step sizes for activations associated with each token
- Per-Channel Quantization
- Group-Wise Quantization



1	0	1	0
0	0	-1	1
0	1	0	-1
1	0	1	1

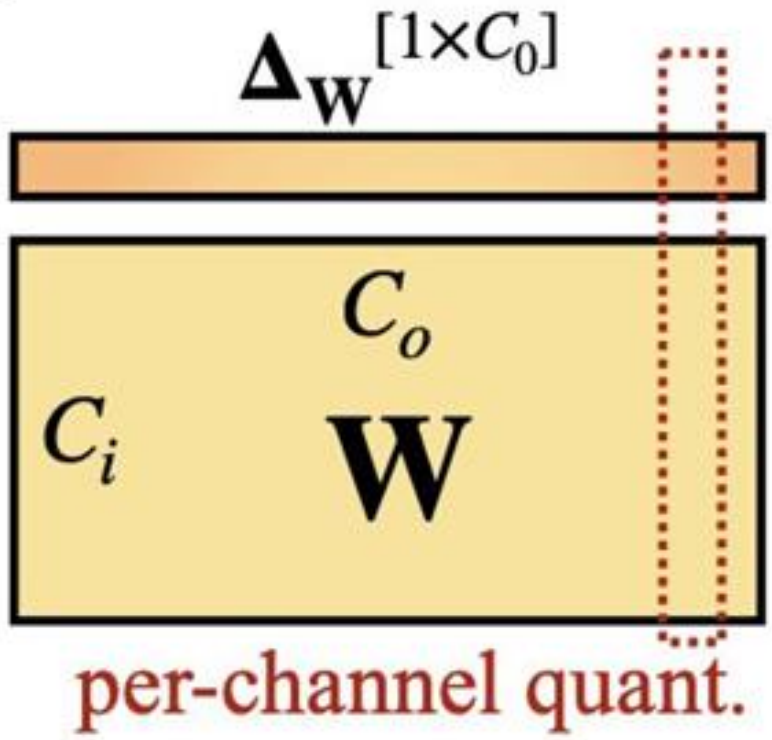
Quantized

2.09	0	2.09	0
0	0	-2.12	2.12
0	1.92	0	-1.92
1.87	0	1.87	1.87

Reconstructed

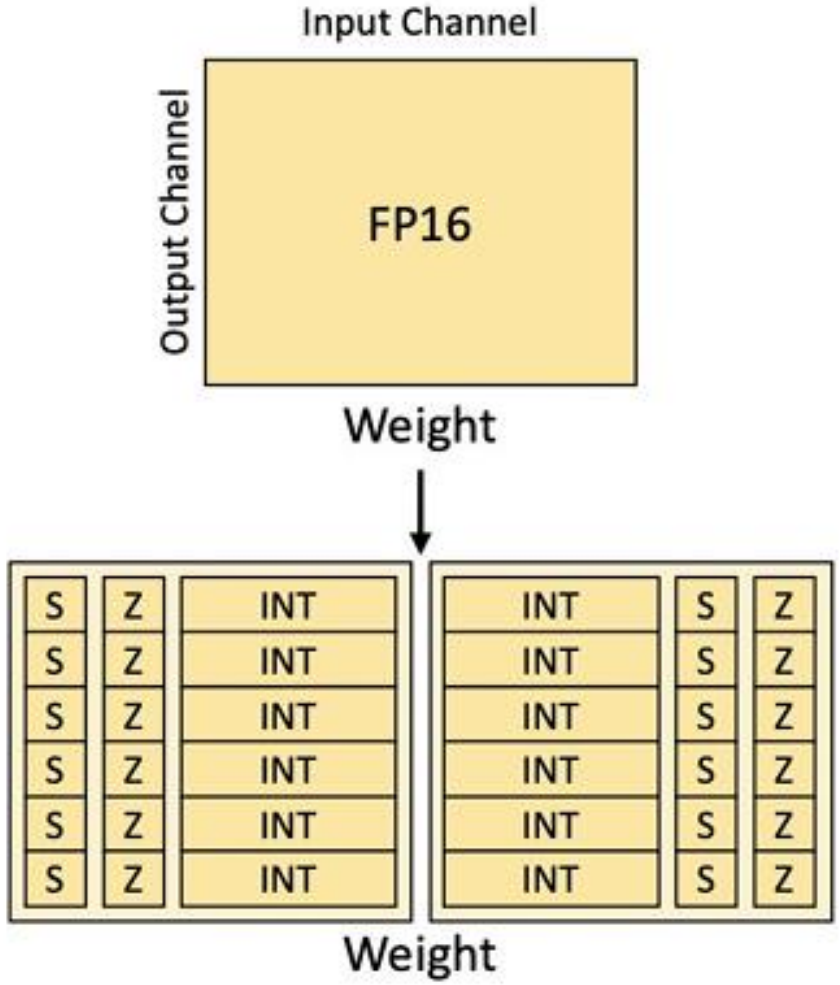
Quantization Granularity

- Per-Tensor Quantization
- Per-Token Quantization
- Per-Channel Quantization
  - Uses different quantization step sizes for activations associated with each output channel of weights
- Group-Wise Quantization



Quantization Granularity

- Per-Tensor Quantization
- Per-Token Quantization
- Per-Channel Quantization
- Group-Wise Quantization
  - Different quantization steps for different channel groups



ZeroQuant: Fine-Grained Quantization

- Weight Quantization: Group-Wise

- Activation: Token-wise Quantization
  - Fine-grained
  - Dynamically calculate the min/max range

ZeroQuant: LKD

- Layer-by-layer distillation (LKD)
  - Teacher model: Original (i.e., unquantized) version
    - Use the output of the L_{k-1} as the input of L_k
  - Student model: Quantized version

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

ZeroQuant: Optimized Kernels

- Optimized Transformer Kernels
  - CUTLASS INT8 GeMM
  - Fusing Token-wise Activation Quantization + GeMM

Evaluation Results

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 [57] (QAT)$^{+}$	—	83.91	83.83	—	—	—	—	92.83	—	—	—
W8A8 [78] (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 [7] (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

https://proceedings.neurips.cc/paper_files/paper/2022/file/adf7fa39d65e2983d724ff7da57f00ac-Paper-Conference.pdf

Evaluation Results

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 [57] (QAT)+	—	83.91	83.83	—	—	—	—	92.83	—	—	—
W8A8 [78] (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 [7] (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

https://proceedings.neurips.cc/paper_files/paper/2022/file/adf7fa39d65e2983d724ff7da57f00ac-Paper-Conference.pdf

Evaluation Results

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 [57] (QAT)+	—	83.91	83.83	—	—	—	—	92.83	—	—	—
W8A8 [78] (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 [7] (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

https://proceedings.neurips.cc/paper_files/paper/2022/file/adf7fa39d65e2983d724ff7da57f00ac-Paper-Conference.pdf

Evaluation Results

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 [57] (QAT)+	—	83.91	83.83	—	—	—	—	92.83	—	—	—
W8A8 [78] (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 [7] (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

https://proceedings.neurips.cc/paper_files/paper/2022/file/adf7fa39d65e2983d724ff7da57f00ac-Paper-Conference.pdf

Evaluation Results

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 [57] (QAT)+	—	83.91	83.83	—	—	—	—	92.83	—	—	—
W8A8 [78] (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 [7] (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

https://proceedings.neurips.cc/paper_files/paper/2022/file/adf7fa39d65e2983d724ff7da57f00ac-Paper-Conference.pdf

Evaluation Results: Inference Speed

Seq Len		128								256							
BS	Precision	1	2	4	8	16	16	64	128	1	2	4	8	16	16	64	128
$BERT_{base}$	W16A16	2.45	3.22	3.85	5.51	9.96	17.93	34.25	67.08	3.13	4.05	5.70	10.55	19.27	36.69	71.75	140.0
	W8A8	1.08	1.16	1.42	1.76	2.58	3.90	6.74	12.92	1.22	1.44	2.08	2.88	4.10	7.80	14.66	28.13
	Speedup	2.27	2.78	2.71	3.13	3.86	4.60	5.08	5.19	2.57	2.81	2.74	3.66	4.70	4.70	4.89	4.98
$BERT_{large}$	W16A16	5.45	6.38	8.73	13.88	26.34	48.59	92.49	183.4	6.39	8.94	14.66	27.99	51.94	98.78	195.9	384.5
	W8A8	2.08	2.58	2.84	3.79	6.21	10.28	18.86	36.62	2.55	3.36	4.16	6.88	11.61	21.20	41.24	79.90
	Speedup	2.62	2.47	3.07	3.66	4.24	4.73	4.90	5.01	2.51	2.66	3.52	4.07	4.47	4.66	4.75	4.81

https://proceedings.neurips.cc/paper_files/paper/2022/file/adf7fa39d65e2983d724ff7da57f00ac-Paper-Conference.pdf

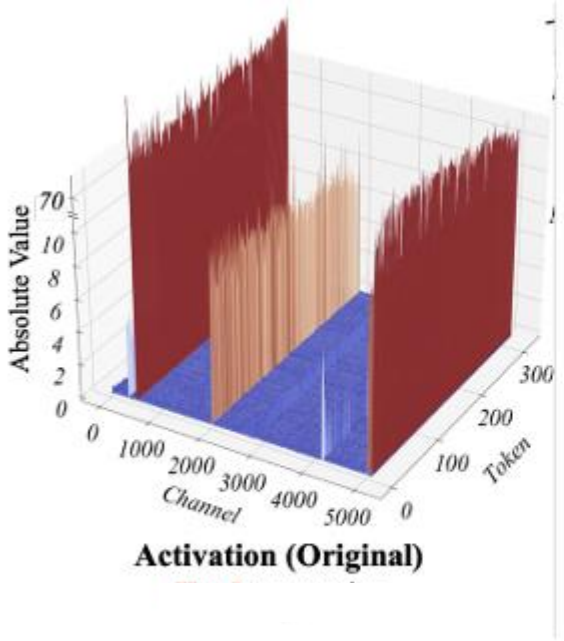
Evaluation Results: Inference Speed

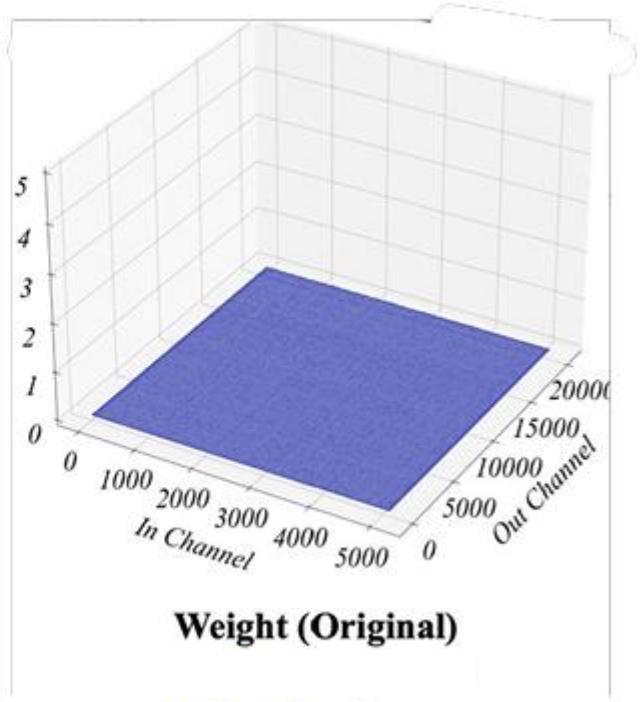
Seq Len		128								256							
BS	Precision	1	2	4	8	16	16	64	128	1	2	4	8	16	16	64	128
$BERT_{base}$	W16A16	2.45	3.22	3.85	5.51	9.96	17.93	34.25	67.08	3.13	4.05	5.70	10.55	19.27	36.69	71.75	140.0
	W8A8	1.08	1.16	1.42	1.76	2.58	3.90	6.74	12.92	1.22	1.44	2.08	2.88	4.10	7.80	14.66	28.13
	Speedup	2.27	2.78	2.71	3.13	3.86	4.60	5.08	5.19	2.57	2.81	2.74	3.66	4.70	4.70	4.89	4.98
$BERT_{large}$	W16A16	5.45	6.38	8.73	13.88	26.34	48.59	92.49	183.4	6.39	8.94	14.66	27.99	51.94	98.78	195.9	384.5
	W8A8	2.08	2.58	2.84	3.79	6.21	10.28	18.86	36.62	2.55	3.36	4.16	6.88	11.61	21.20	41.24	79.90
	Speedup	2.62	2.47	3.07	3.66	4.24	4.73	4.90	5.01	2.51	2.66	3.52	4.07	4.47	4.66	4.75	4.81

https://proceedings.neurips.cc/paper_files/paper/2022/file/adf7fa39d65e2983d724ff7da57f00ac-Paper-Conference.pdf

Limitations

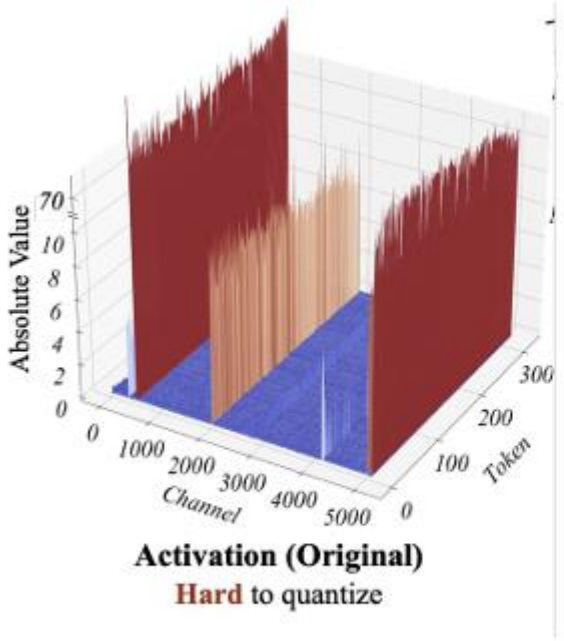
More Observations

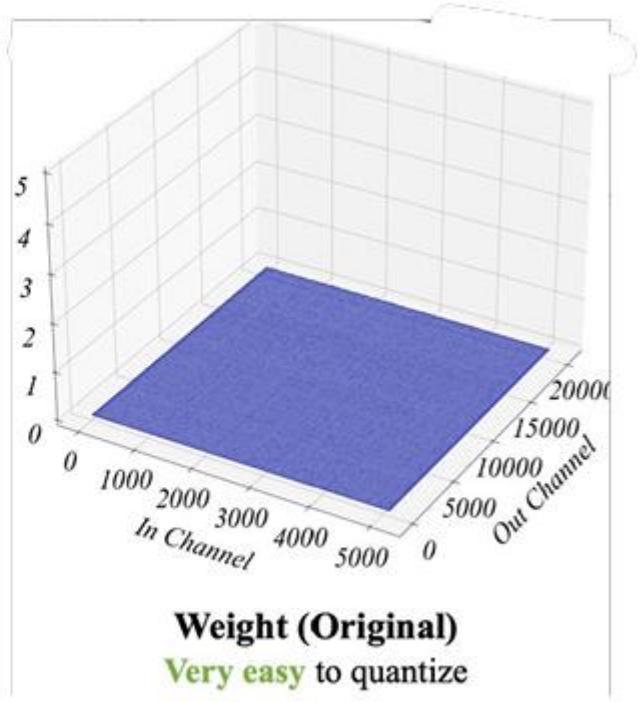




- What is the difference between the two figures?
  - 1 has lots of outliers
  - 3 channels are much higher in value than the surrounding channels
  - Range is 0-70
  - At the same time, 2 is pretty flat

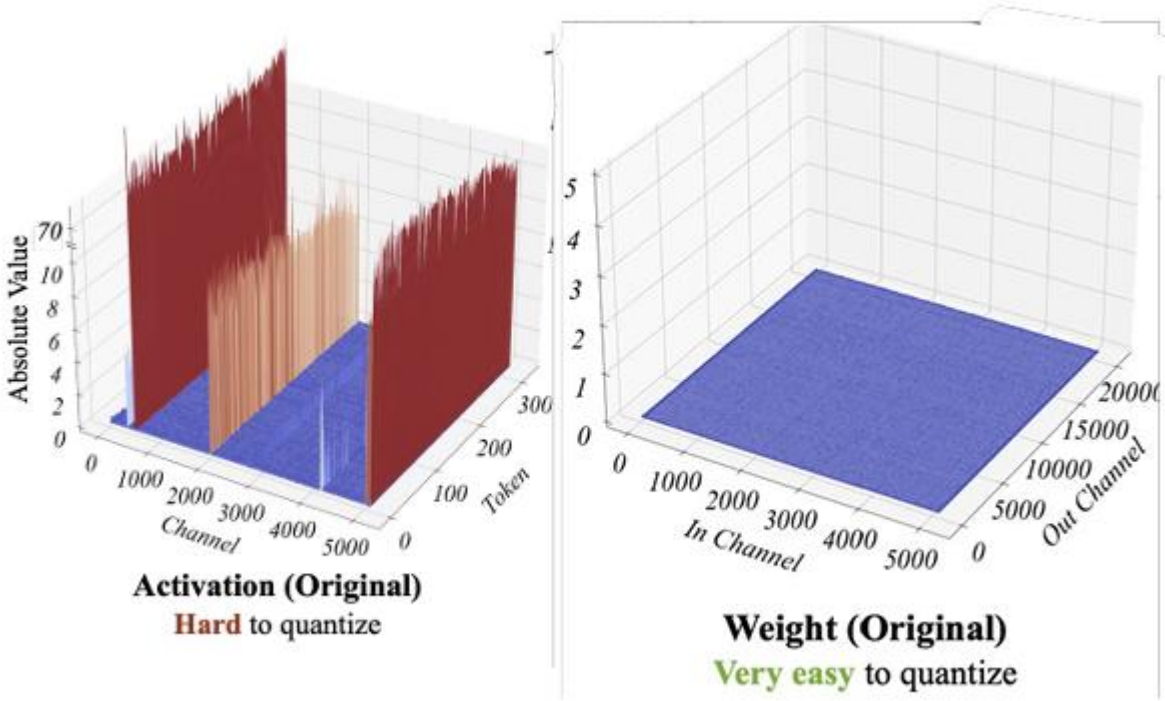
More Observations



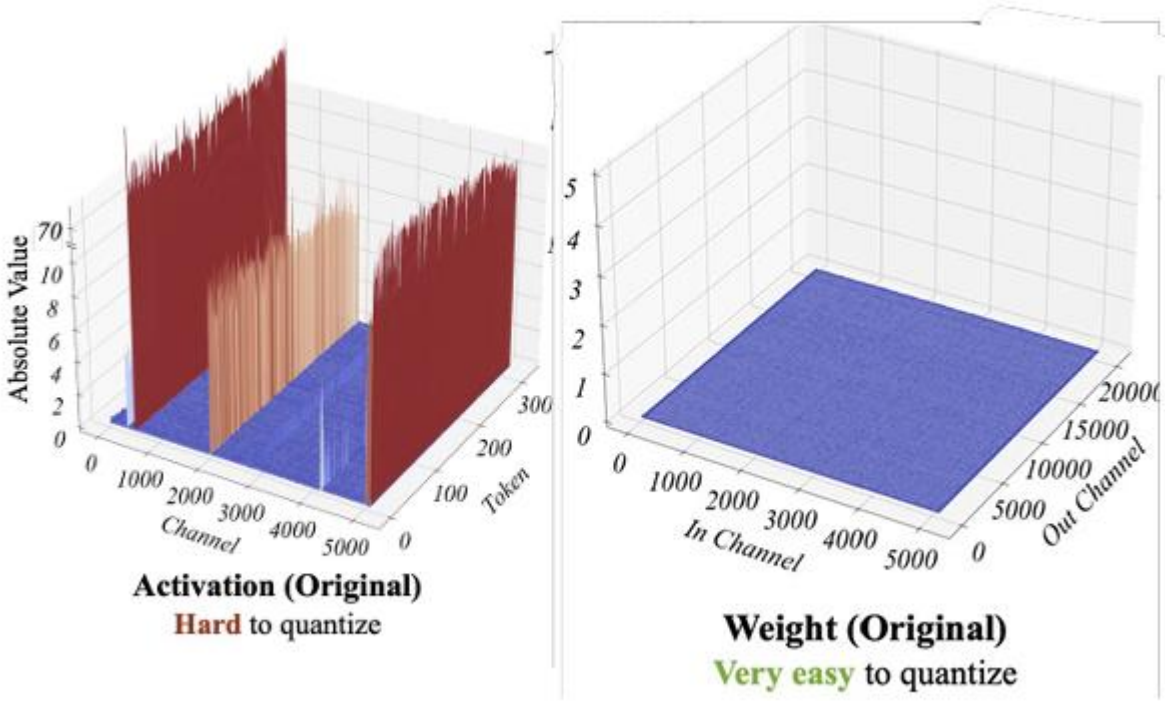


- What is the difference between the two figures?
  - 1 has lots of outliers
  - 3 channels are much higher in value than the surrounding channels
  - Range is 0-70
  - At the same time, 2 is pretty flat

SmoothQuant

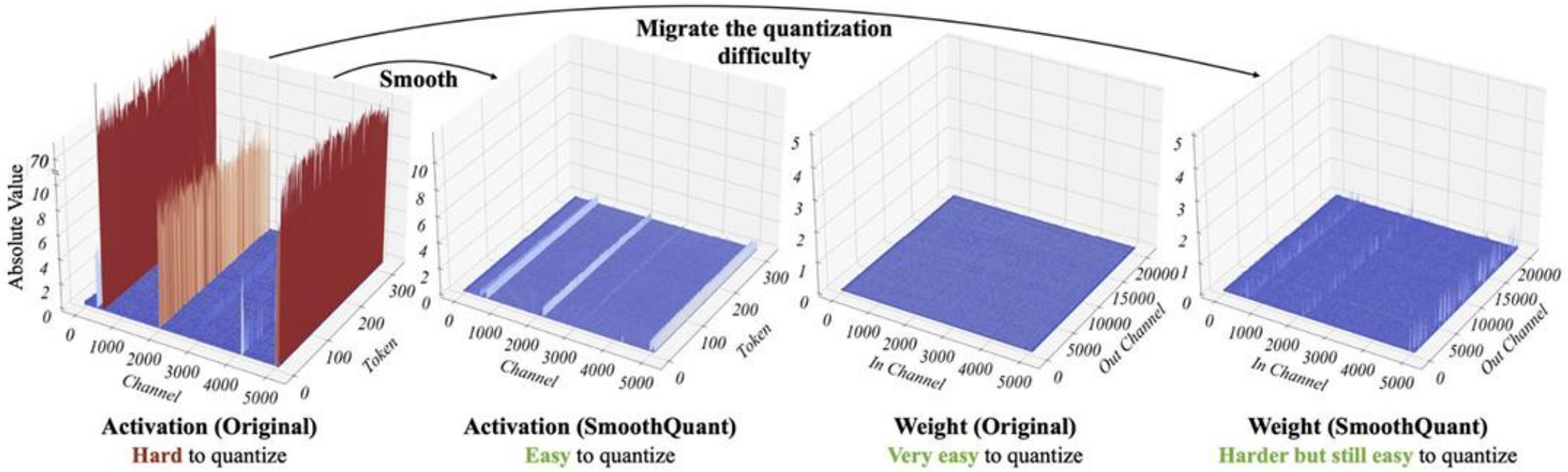


SmoothQuant



$$\mathbf{Y} = \mathbf{XW} = (\mathbf{0.01X})(\mathbf{100W})$$

SmoothQuant



- **Key idea: Migrating the quantization difficulty from activation to weight**

SmoothQuant: Per-Channel Smoothing Factor

$$\mathbf{Y} = (\mathbf{X}\text{diag}(\mathbf{s})^{-1}) \cdot (\text{diag}(\mathbf{s})\mathbf{W}) = \hat{\mathbf{X}}\hat{\mathbf{W}}$$

- Push all quantization difficulty from activations to weights (all channels have same maximum magnitude):

$$\mathbf{s}_j = \max(|\mathbf{X}_j|), j = 1, 2, ..., C_i,$$

- Push all quantization difficulty from weights to activations:

$$\mathbf{s}_j = 1/\max(|\mathbf{W}_j|)$$

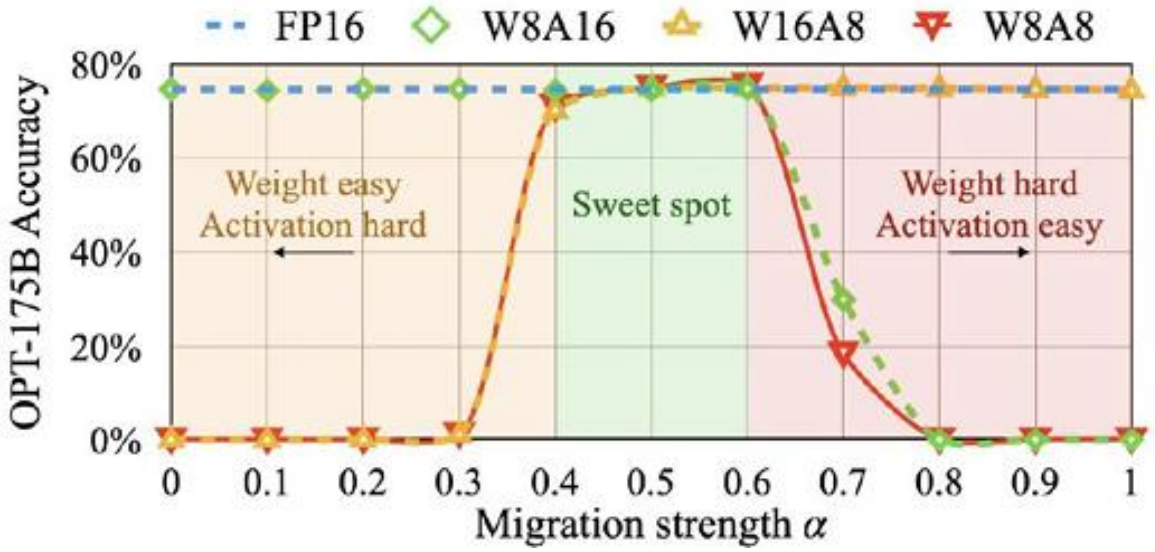
- Share difficulty according to α:

$$\mathbf{s}_j = \max(|\mathbf{X}_j|)^{\alpha} / \max(|\mathbf{W}_j|)^{1-\alpha}$$

SmoothQuant: Alpha (α)

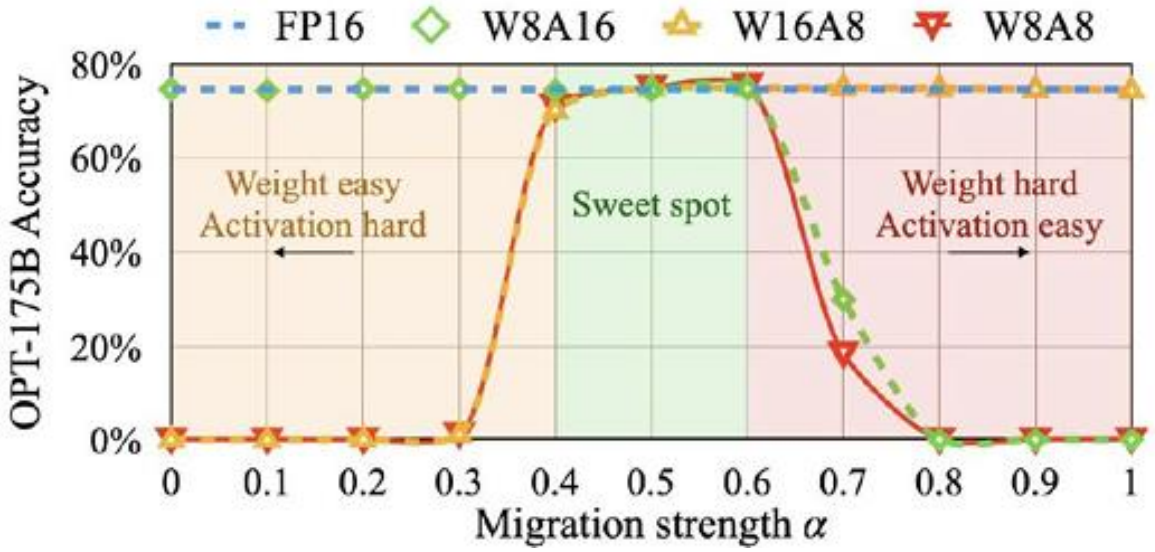
- Hyperparameter which controls the extent to which quantization difficulty is shifted from activations to weights

- α is b/w 0.4 to 0.6, though larger models or models with more significant activation outliers may require higher values.



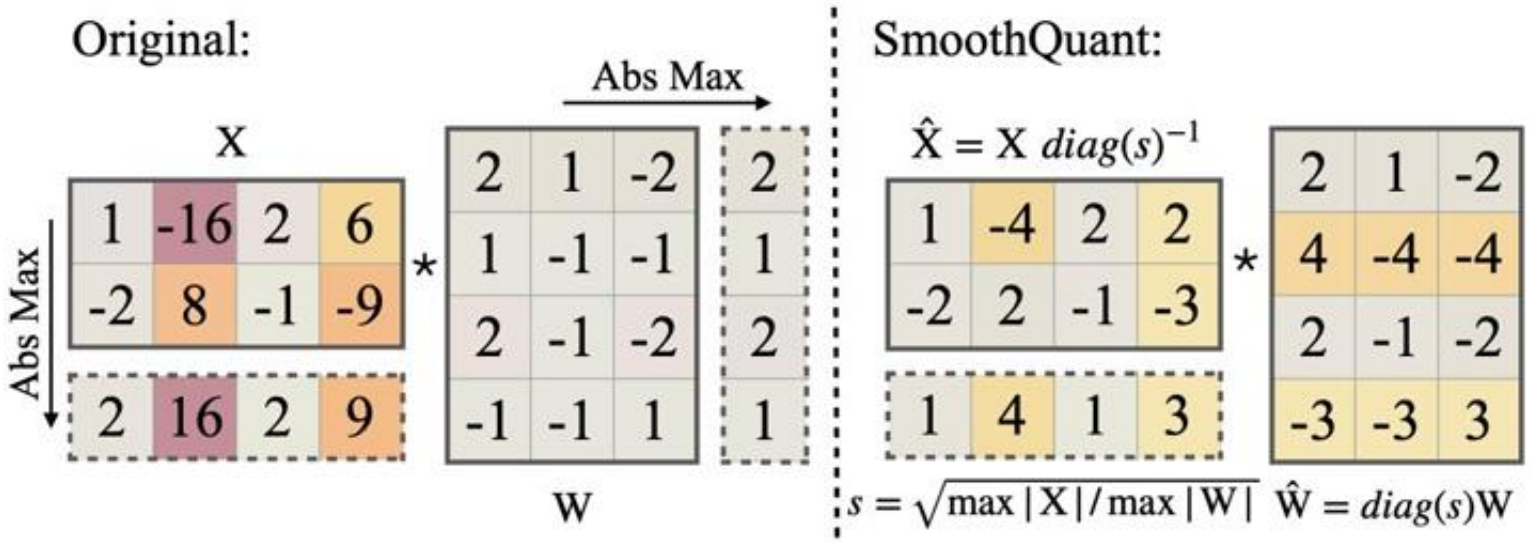
Choosing α

- Case-by-case decision
- If α is too large, weights will be hard to quantize. If too small, activations will be hard to quantize.
- Goal: Make activations and weights both easy to quantize.



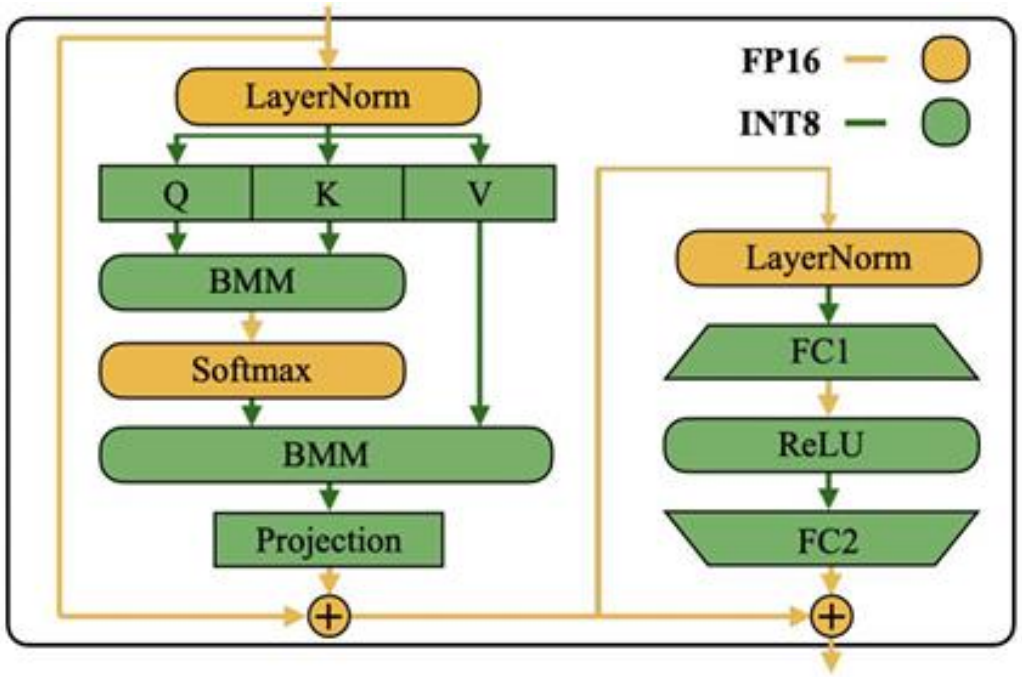
Example of SmoothQuant

- 1) Applying Smoothing Factor
- 2) Quantize (constant step size)



SmoothQuant Hardware Efficiency

- Applying SmoothQuant to transfer blocks
  - Linear layers take up most of the parameter and computation
  - Smoothing factor can be fused into previous layers' parameters offline
  - All linear layers are quantized with W8A8, as well as BMM operators in attention computation



Evaluation: Baselines

- W8A8 is the naïve implementation
- LLM.int8() keeps outliers in FP16
- Outlier suppression uses token-wise clipping

Method	Weight	Activation
W8A8	per-tensor	per-tensor dynamic
ZeroQuant	group-wise	per-token dynamic
`LLM.int8()`	per-channel	per-token dynamic+FP16
Outlier Suppression	per-tensor	per-tensor static

Evaluation: SmoothQuant O1 to O3

- Gradually aggressive and efficient (lower latency) quantization levels

Method	Weight	Activation
SmoothQuant-O1	per-tensor	per-token dynamic
SmoothQuant-O2	per-tensor	per-tensor dynamic
SmoothQuant-O3	per-tensor	per-tensor static

Evaluation: OPT-175B

OPT-175B	LAMBADA	HellaSwag	PIQA	WinoGrande	OpenBookQA	RTE	COPA	**Average↑**	**WikiText↓**
FP16	74.7%	59.3%	79.7%	72.6%	34.0%	59.9%	88.0%	66.9%	10.99
W8A8	0.0%	25.6%	53.4%	50.3%	14.0%	49.5%	56.0%	35.5%	93080
ZeroQuant	0.0%*	26.0%	51.7%	49.3%	17.8%	50.9%	55.0%	35.8%	84648
`LLM.int8()`	74.7%	59.2%	79.7%	72.1%	34.2%	60.3%	87.0%	66.7%	11.10
Outlier Suppression	0.00%	25.8%	52.5%	48.6%	16.6%	53.4%	55.0%	36.0%	96151
SmoothQuant-O1	74.7%	59.2%	79.7%	71.2%	33.4%	58.1%	89.0%	66.5%	11.11
SmoothQuant-O2	75.0%	59.0%	79.2%	71.2%	33.0%	59.6%	88.0%	66.4%	11.14
SmoothQuant-O3	74.6%	58.9%	79.7%	71.2%	33.4%	59.9%	90.0%	66.8%	11.17

Evaluation: Results on Different LLMs

Method	OPT-175B	BLOOM-176B	GLM-130B*
FP16	71.6%	68.2%	73.8%
W8A8	32.3%	64.2%	26.9%
ZeroQuant	31.7%	67.4%	26.7%
`LLM.int8()`	71.4%	68.0%	73.8%
Outlier Suppression	31.7%	54.1%	63.5%
SmoothQuant-O1	**71.2%**	68.3%	**73.7%**
SmoothQuant-O2	71.1%	**68.4%**	72.5%
SmoothQuant-O3	71.1%	67.4%	72.8%

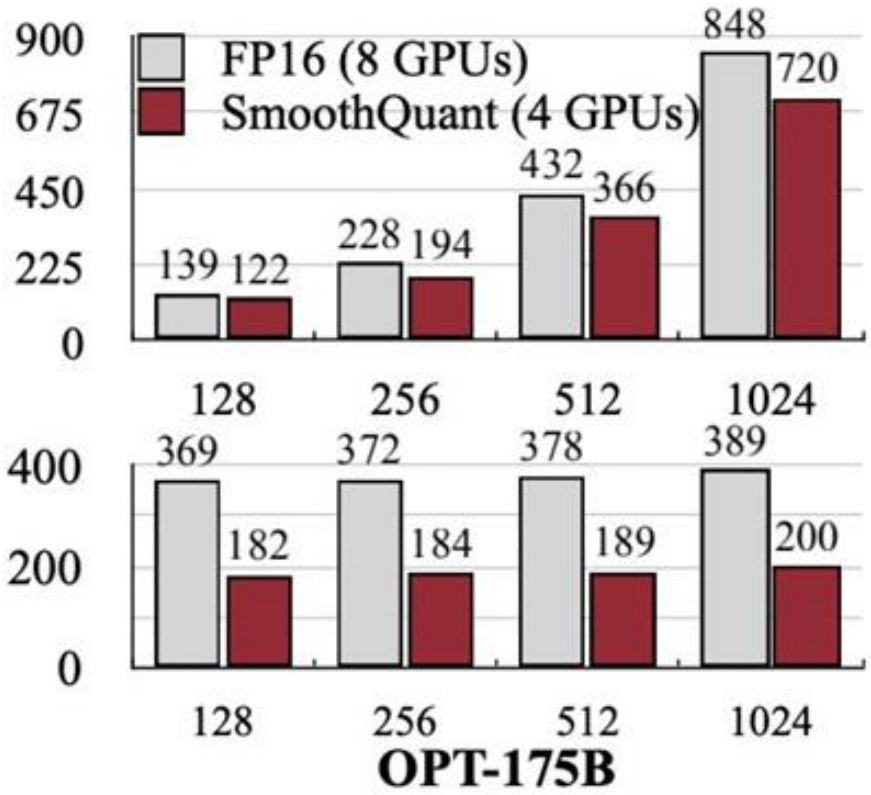
Evaluation: Near Lossless W8A8 Quantization for LLaMA

Wiki PPL↓	7B	13B	30B	65B
FP16	11.51	10.05	7.53	6.17
W8A8 SmoothQuant	11.56	10.08	7.56	6.20

Similar!

Hardware Efficiency

- Similar or faster latency with half #GPUs



Limitations

Questions?