RSQ: Learning from Important Tokens Leads to Better Quantized LLMs

Yi-Lin Sung

RSQ (Rotate, Scale, then Quantize)

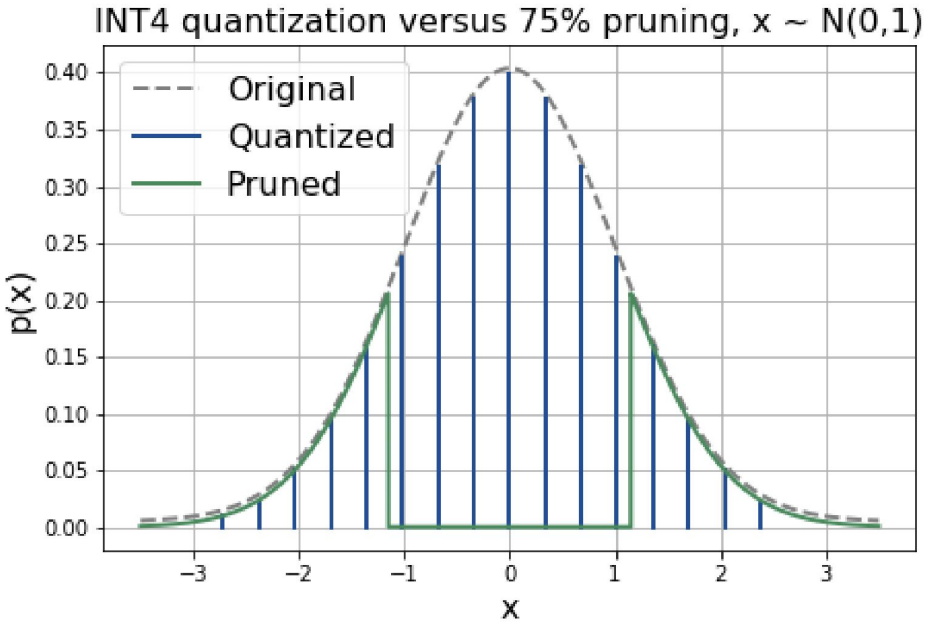
- **Rotate**: rotate the model to **reduce outliers**
- **Scale**: scale the tokens based on importance
- **Quantize**: quantize the model using layer-wise quantization method (GPTQ) with scaled tokens

RSQ (Rotate, Scale, then Quantize)

- **Rotate**: rotate the model to **reduce outliers**
- **Scale**: scale the tokens based on importance
- **Quantize**: quantize the model using layer-wise quantization method (GPTQ) with scaled tokens

Pruning and quantization

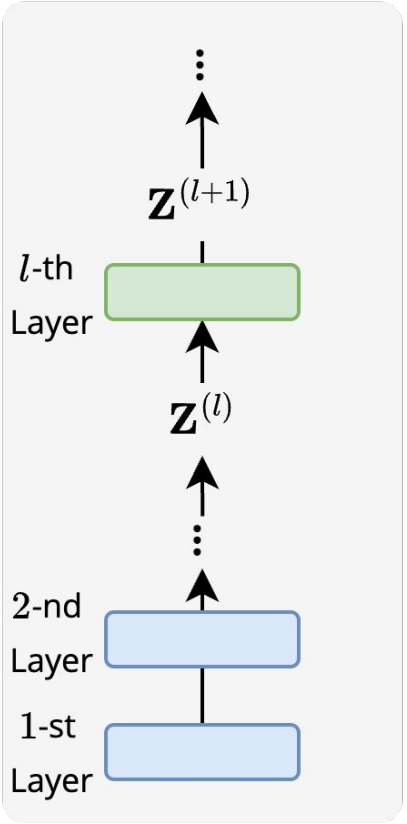
- Pruning: **remove** less important weight parameters
- Quantization: use **fewer bits** to represent weight parameters



Layer-wise quantization

- Calibrated on a small dataset

$$\text{argmin}_{\hat{\mathbf{W}}}||\mathbf{WX} - \hat{\mathbf{W}}\mathbf{X}||_2^2, \text{s.t } \mathbf{W}_p + \mathrm{e}_p^T \cdot \delta\mathbf{W} = \text{quant}(\mathbf{W}_p)$$



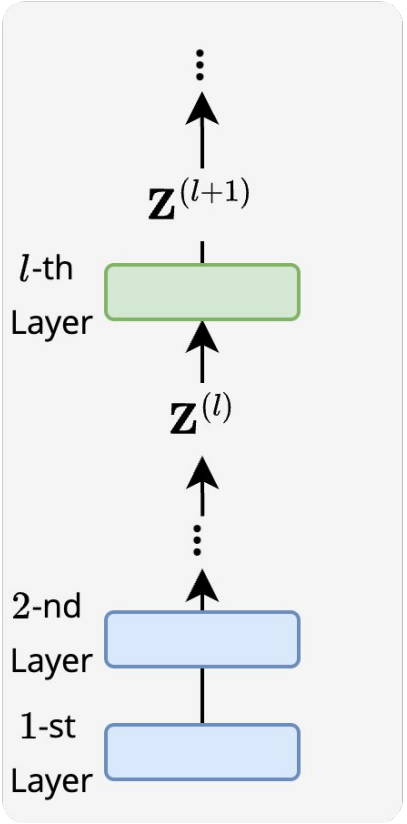
Layer-wise quantization

- Calibrated on a small dataset

$$\text{argmin}_{\hat{\mathbf{W}}} ||\mathbf{WX} - \hat{\mathbf{W}}\mathbf{X}||_2^2, \text{s.t } \mathbf{W}_p + \mathbf{e}_p^T \cdot \delta\mathbf{W} = \text{quant}(\mathbf{W}_p)$$

Layer reconstruction loss

Quantize one vector



Closed-form solution

$$w_q = \text{argmin}_{w_q} \frac{(\text{quant}(w_q) - w_q)^2}{[\mathbf{H}_F^{-1}]_{qq}}, \quad \boldsymbol{\delta}_F = -\frac{w_q - \text{quant}(w_q)}{[\mathbf{H}_F^{-1}]_{qq}} \cdot (\mathbf{H}_F^{-1})_{:,q}.$$

$$\mathbf{H}_F = 2\mathbf{X}_F\mathbf{X}_F^\top$$

RSQ (Rotate, Scale, then Quantize)

- **Rotate**: rotate the model to **reduce outliers**
- **Scale**: scale the tokens based on importance
- **Quantize**: quantize the model using layer-wise quantization method (GPTQ) with scaled tokens

Layer-wise quantization

- Pros
  - Fast and efficient
  - The GPU memory cost does not scale with number of layers
  - Strong performance with a small calibration dataset
- Cons
  - Fail to consider importance across layers and modules

Layer-wise quantization

- Pros
  - Fast and efficient
  - The GPU memory cost does not scale with number of layers
  - Strong performance with a small calibration dataset
- Cons
  - Fail to consider importance across layers and modules
  - Minimizing layer reconstruction loss may not be optimal

Layer-wise quantization

- Pros
  - Fast and efficient
  - The GPU memory cost does not scale with number of layers
  - Strong performance with a small calibration dataset
- Cons
  - Fail to consider importance across layers and modules
  - Minimizing layer reconstruction loss may not be optimal
  - Does not scale well to a larger calibration dataset

Layer-wise quantization

- Pros
  - Fast and efficient
  - The GPU memory cost does not scale with number of layers
  - Strong performance with a small calibration dataset
- Cons
  - Fail to consider importance across layers and modules
  - **Minimizing layer reconstruction loss may not be optimal**
  - Does not scale well to a larger calibration dataset

Standard layer-wise quantization

$$\text{argmin}_{\hat{\mathbf{W}}} ||\mathbf{WX} - \hat{\mathbf{W}}\mathbf{X}||_2^2, \text{s.t} \quad \mathbf{W}_p + \mathbf{e}_p^T \cdot \delta\mathbf{W} = \text{quant}(\mathbf{W}_p)$$

Layer reconstruction loss

Quantize one weight

Standard layer-wise quantization

$$||\mathbf{WX} - \hat{\mathbf{W}}\mathbf{X}||_2^2$$

Standard layer-wise quantization

$$||\mathbf{W}\mathbf{X} - \hat{\mathbf{W}}\mathbf{X}||_2^2 = \sum_i ||\mathbf{W}\mathbf{x}_i - \hat{\mathbf{W}}\mathbf{x}_i||_2^2$$

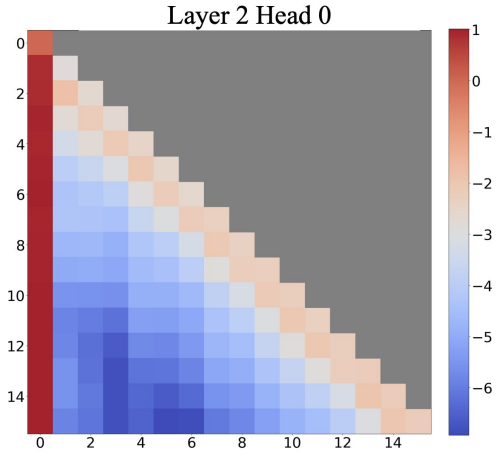
Standard layer-wise quantization

$$||\mathbf{W}\mathbf{X} - \hat{\mathbf{W}}\mathbf{X}||_2^2 = \sum_i ||\mathbf{W}\mathbf{x}_i - \hat{\mathbf{W}}\mathbf{x}_i||_2^2$$

Uniform across tokens!

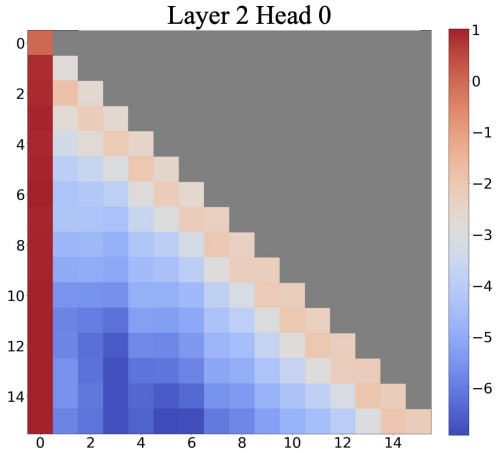
Tokens are not equally important

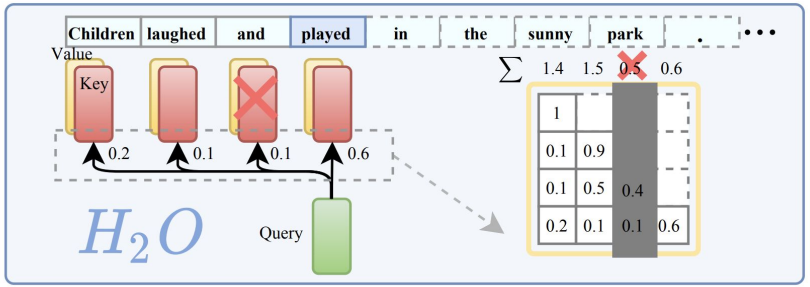
- AttnSink: query attends on initial and last tokens more



Tokens are not equally important

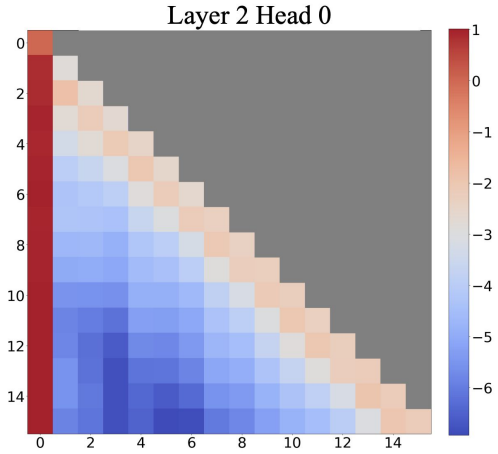
- AttnSink: query attends on initial and last tokens more
- H2O: not all KV caches are equal important

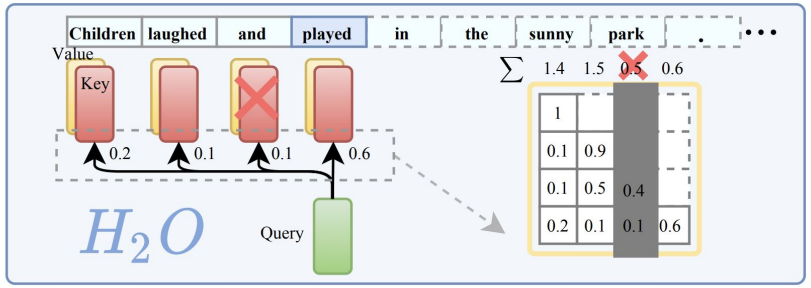


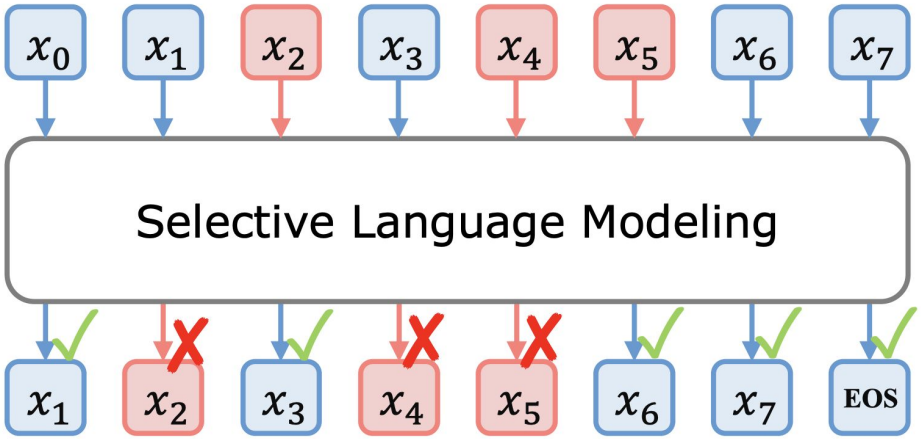


Tokens are not equally important

- AttnSink: query attends on initial and last tokens more
- H2O: not all KV caches are equal important
- RHO-1: train LLMs with only a subset of tokens







Uniform loss is optimal?

$$||\mathbf{W}\mathbf{X} - \hat{\mathbf{W}}\mathbf{X}||_2^2 = \sum_i ||\mathbf{W}\mathbf{x}_i - \hat{\mathbf{W}}\mathbf{x}_i||_2^2$$

Uniform loss is optimal?

$$||\mathbf{WX} - \hat{\mathbf{W}}\mathbf{X}||_2^2 = \sum_i ||\mathbf{W}\mathbf{x}_i - \hat{\mathbf{W}}\mathbf{x}_i||_2^2$$

Table 1. Evaluation results of quantizing LLaMA-3-8B-Instruct with WikiText-2 using different subsets of tokens.

Used Token IDs	All: 1 - 4096	1 - 1024	1025 - 2048	2049 - 3072	3073 - 4096
Wiki PPL ↓	$9.51_{.11}$	$\mathbf{9.27}_{.05}$	$10.26_{.12}$	$10.16_{.08}$	$10.25_{.01}$
Avg Acc (%) ↑	$63.8_{0.5}$	$\mathbf{64.5}_{0.4}$	$61.7_{0.7}$	$61.4_{0.4}$	$61.3_{0.5}$



Our objective

$$\|(\mathbf{W}\mathbf{X} - \tilde{\mathbf{W}}\mathbf{X})\mathbf{R}\|_2^2 = \sum_{i=0}^{T} \|r_i(\mathbf{W}\mathbf{x}_i - \tilde{\mathbf{W}}\mathbf{x}_i)\|_2^2$$

Our objective

$$\|(\mathbf{W}\mathbf{X} - \tilde{\mathbf{W}}\mathbf{X})\mathbf{R}\|_2^2 = \sum_{i=0}^{T} \|r_i(\mathbf{W}\mathbf{x}_i - \tilde{\mathbf{W}}\mathbf{x}_i)\|_2^2$$

$$w_q = \text{argmin}_{w_q} \frac{(\text{quant}(w_q) - w_q)^2}{[\mathbf{H}_F^{-1}]_{qq}}, \quad \boldsymbol{\delta}_F = -\frac{w_q - \text{quant}(w_q)}{[\mathbf{H}_F^{-1}]_{qq}} \cdot (\mathbf{H}_F^{-1})_{:,q}.$$

$$\mathbf{H}_{\text{RSQ}} = 2\mathbf{X}\mathbf{R}^2\mathbf{X}^\top$$

Our objective

$$\|(\mathbf{W}\mathbf{X} - \tilde{\mathbf{W}}\mathbf{X})\mathbf{R}\|_2^2 = \sum_{i=0}^{T} \|r_i(\mathbf{W}\mathbf{x}_i - \tilde{\mathbf{W}}\mathbf{x}_i)\|_2^2$$

1. Heuristic Approaches:
    a. First-N: First N tokens
    b. First&Last-N: First N/2 tokens + Last N/2 tokens

Our objective

$$\|(\mathbf{W}\mathbf{X} - \tilde{\mathbf{W}}\mathbf{X})\mathbf{R}\|_2^2 = \sum_{i=0}^{T} \|r_i(\mathbf{W}\mathbf{x}_i - \tilde{\mathbf{W}}\mathbf{x}_i)\|_2^2$$

1. Heuristic Approaches:
   a. First-N: First N tokens
   b. First&Last-N: First N/2 tokens + Last N/2 tokens
2. Dynamic Approaches:
   a. TokenFreq: token occurring frequency

Our objective

$$\|(\mathbf{W}\mathbf{X} - \tilde{\mathbf{W}}\mathbf{X})\mathbf{R}\|_2^2 = \sum_{i=0}^{T} \|r_i(\mathbf{W}\mathbf{x}_i - \tilde{\mathbf{W}}\mathbf{x}_i)\|_2^2$$

1. Heuristic Approaches:
   a. First-N: First N tokens
   b. First&Last-N: First N/2 tokens + Last N/2 tokens
2. Dynamic Approaches:
   a. TokenFreq: token occurring frequency
   b. ActNorm: norm of the token

Our objective

$$\|(\mathbf{W}\mathbf{X} - \tilde{\mathbf{W}}\mathbf{X})\mathbf{R}\|_2^2 = \sum_{i=0}^{T} \|r_i(\mathbf{W}\mathbf{x}_i - \tilde{\mathbf{W}}\mathbf{x}_i)\|_2^2$$

1. Heuristic Approaches:
   a. First-N: First N tokens
   b. First&Last-N: First N/2 tokens + Last N/2 tokens
2. Dynamic Approaches:
   a. TokenFreq: token occurring frequency
   b. ActNorm: norm of the token
   c. ActDiff: feature change of the token after applying one layer

Our objective

$$\|(\mathbf{WX} - \tilde{\mathbf{W}}\mathbf{X})\mathbf{R}\|_2^2 = \sum_{i=0}^{T} \|r_i(\mathbf{W}\mathbf{x}_i - \tilde{\mathbf{W}}\mathbf{x}_i)\|_2^2$$

1. Heuristic Approaches:
   a. First-N: First N tokens
   b. First&Last-N: First N/2 tokens + Last N/2 tokens
2. Dynamic Approaches:
   a. TokenFreq: token occurring frequency
   b. ActNorm: norm of the token
   c. ActDiff: feature change of the token after applying one layer
   d. TokenSim: token's similarity to other tokens

Our objective

$$\|(\mathbf{W}\mathbf{X} - \tilde{\mathbf{W}}\mathbf{X})\mathbf{R}\|_2^2 = \sum_{i=0}^{T} \|r_i(\mathbf{W}\mathbf{x}_i - \tilde{\mathbf{W}}\mathbf{x}_i)\|_2^2$$

1. Heuristic Approaches:
   a. First-N: First N tokens
   b. First&Last-N: First N/2 tokens + Last N/2 tokens
2. Dynamic Approaches:
   a. TokenFreq: token occurring frequency
   b. ActNorm: norm of the token
   c. ActDiff: feature change of the token after applying one layer
   d. TokenSim: token's similarity to other tokens
   e. AttnCon: attention concentration on the token

Our objective

$$\|(\mathbf{W}\mathbf{X} - \tilde{\mathbf{W}}\mathbf{X})\mathbf{R}\|_2^2 = \sum_{i=0}^{T} \|r_i(\mathbf{W}\mathbf{x}_i - \tilde{\mathbf{W}}\mathbf{x}_i)\|_2^2$$

1. Heuristic Approaches:
   a. First-N: First N tokens
   b. First&Last-N: First N/2 tokens + Last N/2 tokens
2. Dynamic Approaches:
   a. TokenFreq: token occurring frequency
   b. ActNorm: norm of the token
   c. ActDiff: feature change of the token after applying one layer
   d. TokenSim: token's similarity to other tokens
   e. AttnCon: attention concentration on the token

RSQ (Rotate, Scale, then Quantize)

- **Rotate**: rotate the model to **reduce outliers**
- **Scale**: scale the tokens based on importance
- **Quantize**: quantize the model using layer-wise quantization method (GPTQ) with scaled tokens

Rotation: reduce outliers

$$\mathbf{Y} = \mathbf{W}_2\mathbf{W}_1\mathbf{X}$$

Rotation: reduce outliers

Orthogonal Transformation (Rotate)

$$\mathbf{Y} = \mathbf{W}_2\mathbf{W}_1\mathbf{X}$$

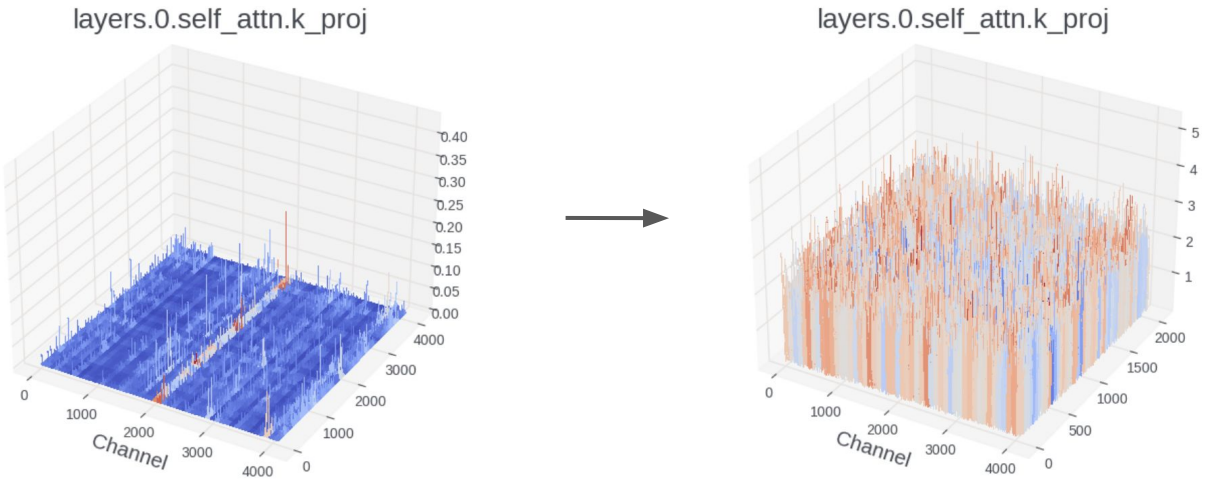
$$\longrightarrow \quad \mathbf{Y} = (\mathbf{W}_2\mathbf{Q}^\top)(\mathbf{Q}\mathbf{W}_1)\mathbf{X}$$

Rotation: reduce outliers

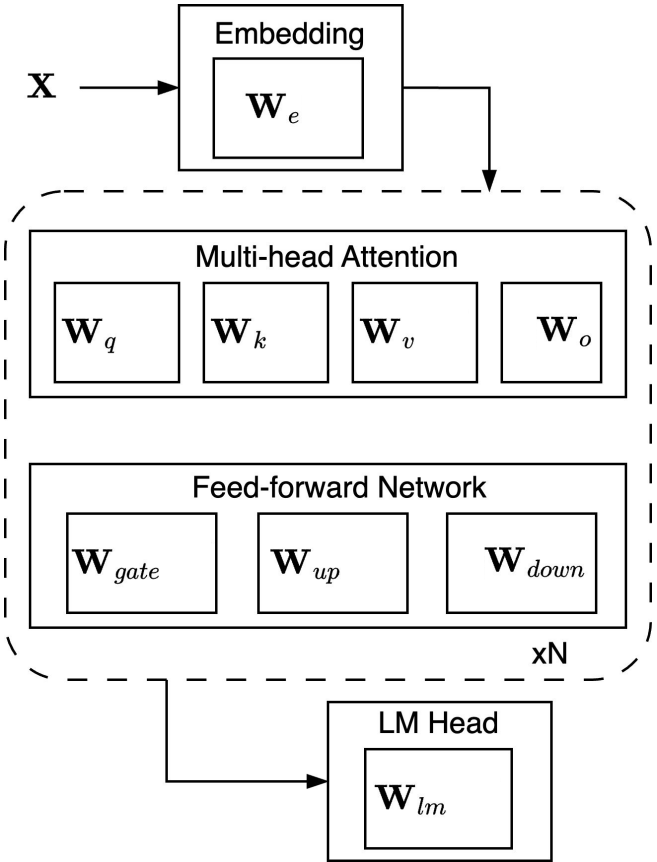
Orthogonal Transformation (Rotate)

$$\mathbf{Y} = \mathbf{W}_2\mathbf{W}_1\mathbf{X}$$

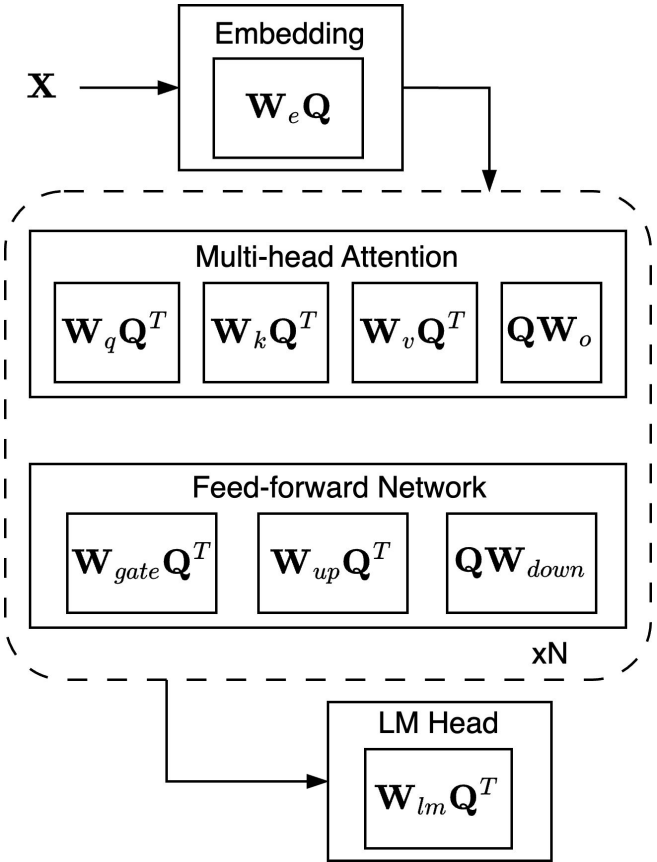
$$\mathbf{Y} = (\mathbf{W}_2\mathbf{Q}^\top)(\mathbf{Q}\mathbf{W}_1)\mathbf{X}$$



Rotation: reduce outliers

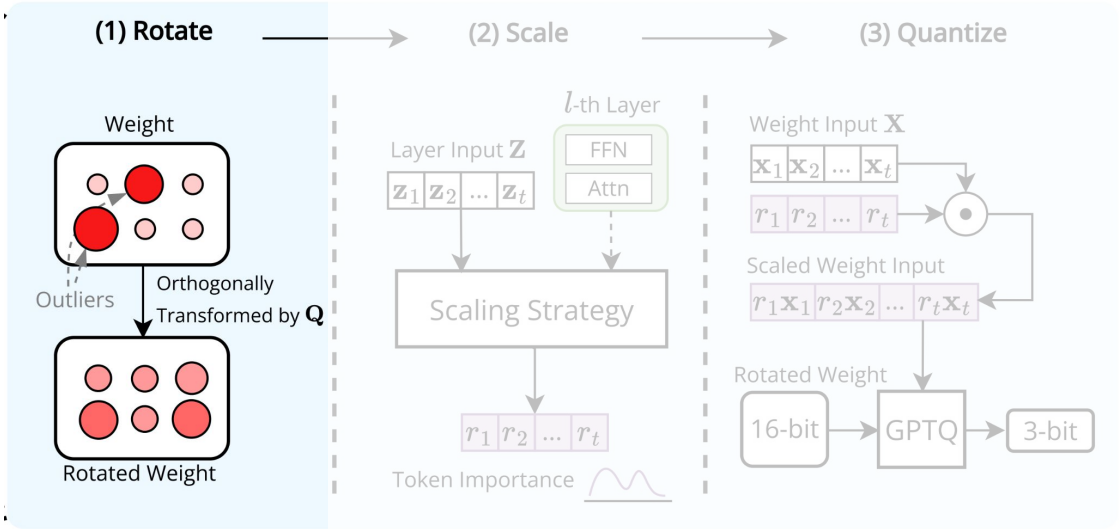


Rotation: reduce outliers



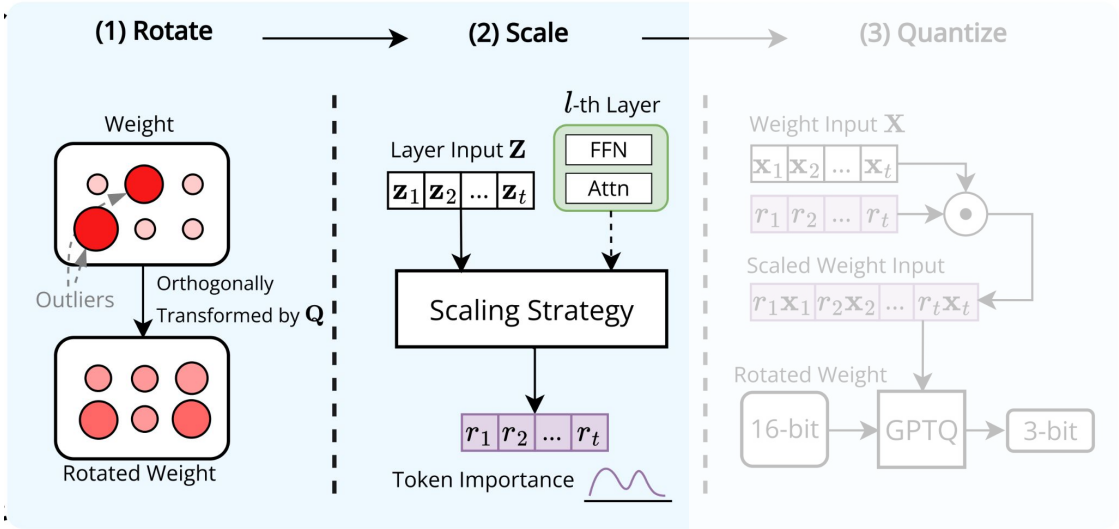
RSQ (Rotate, Scale, then Quantize)

- **Rotate**: rotate the model to **reduce outliers**



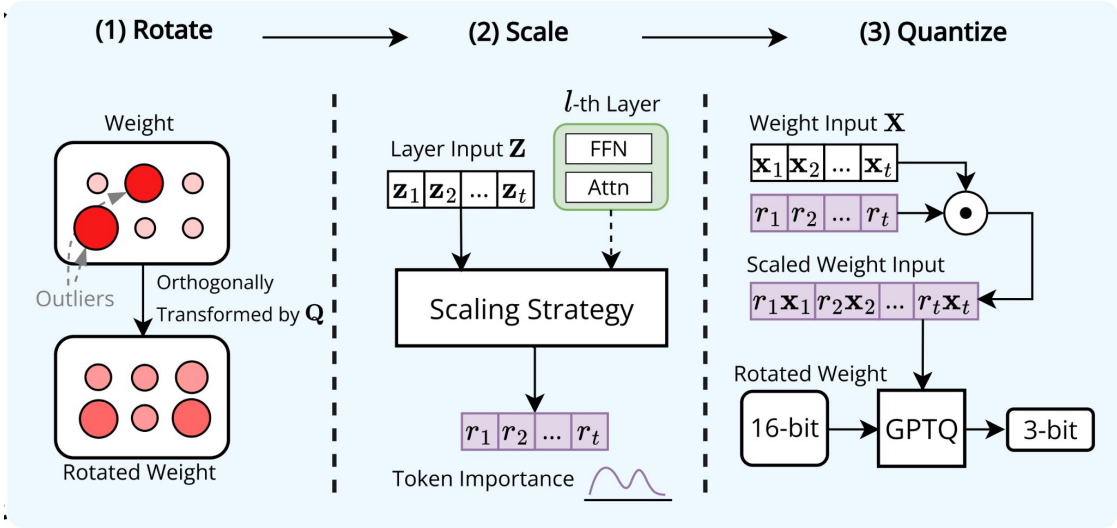
RSQ (Rotate, Scale, then Quantize)

- **Rotate**: rotate the model to **reduce outliers**
- **Scale**: scale the tokens based on importance



RSQ (Rotate, Scale, then Quantize)

- **Rotate**: rotate the model to **reduce outliers**
- **Scale**: scale the tokens based on importance
- **Quantize**: quantize the model using layer-wise quantization method (GPTQ) with scaled tokens



Results

Table 2. Comparison of recent layer-wise quantization approaches with RSQ on multiple downstream tasks. We report perplexity for WikiText and accuracy for all other tasks. The model is quantized to 3-bit. The best-performing method across all quantization approaches is highlighted in bold. We denote the standard deviation across three runs as a subscript.

Method	Wiki	LAMBADA_{openai}	LAMBADA_{std}	WinoGrande	ArcC	ArcE	HellaSwag	PIQA	MMLU	GSM8k	TruthfulQA	Avg
LLaMA3-8B-Instruct												
Full Model	8.311	71.9	65.0	71.7	56.7	79.6	75.8	78.5	65.6	79.9	51.7	69.7
GPTQ	$10.682_{.04}$	$50.7_{2.0}$	$44.6_{1.0}$	$65.7_{0.7}$	$35.8_{1.4}$	$56.9_{2.3}$	$65.8_{0.2}$	$70.4_{1.7}$	$50.8_{0.4}$	$27.5_{1.5}$	$44.8_{1.0}$	$51.3_{0.7}$
QuaRot	$9.517_{.11}$	$68.8_{1.2}$	$59.7_{1.0}$	$70.2_{1.2}$	$49.6_{1.5}$	$74.9_{0.9}$	$71.3_{0.0}$	$76.8_{0.7}$	$57.6_{0.5}$	$61.2_{1.7}$	$48.0_{0.9}$	$63.8_{0.5}$
RSQ	$\mathbf{9.046}_{.01}$	$\mathbf{70.8}_{0.3}$	$\mathbf{62.3}_{0.4}$	$\mathbf{70.6}_{0.7}$	$\mathbf{50.3}_{1.1}$	$\mathbf{76.5}_{1.3}$	$\mathbf{72.1}_{0.2}$	$\mathbf{77.1}_{0.5}$	$\mathbf{60.0}_{0.3}$	$\mathbf{63.4}_{2.4}$	$\mathbf{50.6}_{1.7}$	$\mathbf{65.4}_{0.1}$
Mistral-NeMo-12B												
Full Model	6.095	75.8	68.3	75.1	59.1	80.1	82.2	82.1	68.2	80.8	54.8	72.7
GPTQ	$9.537_{.01}$	$44.6_{0.6}$	$29.8_{4.3}$	$57.3_{1.1}$	$38.8_{0.2}$	$58.0_{1.6}$	$60.5_{2.0}$	$71.2_{0.6}$	$48.0_{1.1}$	$37.6_{1.5}$	$47.0_{1.5}$	$49.3_{0.4}$
QuaRot	$6.782_{.00}$	$\mathbf{75.6}_{0.3}$	$60.9_{5.8}$	$72.8_{0.7}$	$\mathbf{55.8}_{1.5}$	$76.9_{1.1}$	$78.1_{0.3}$	$79.7_{0.4}$	$63.5_{0.0}$	$\mathbf{71.1}_{1.0}$	$52.5_{0.9}$	$68.7_{1.1}$
RSQ	$\mathbf{6.673}_{.01}$	$75.4_{0.3}$	$\mathbf{66.5}_{0.6}$	$\mathbf{73.5}_{0.4}$	$55.7_{0.5}$	$\mathbf{77.2}_{1.1}$	$\mathbf{78.5}_{0.1}$	$\mathbf{80.7}_{0.2}$	$\mathbf{64.3}_{0.3}$	$\mathbf{71.1}_{0.4}$	$\mathbf{52.9}_{0.5}$	$\mathbf{69.6}_{0.2}$
Qwen-2.5-7B-Instruct												
Full Model	5.335	75.2	68.6	72.7	59.0	76.4	85.2	81.0	83.3	84.4	65.5	75.1
GPTQ	$9.577_{.03}$	$53.6_{1.0}$	$48.6_{1.8}$	$62.1_{0.9}$	$47.1_{1.8}$	$66.9_{2.1}$	$72.8_{0.0}$	$73.7_{0.5}$	$60.2_{0.5}$	$40.9_{10.5}$	$53.7_{1.5}$	$58.0_{1.8}$
QuaRot	$8.053_{.02}$	$67.9_{0.3}$	$62.1_{0.4}$	$68.3_{1.2}$	$\mathbf{54.0}_{1.6}$	$\mathbf{78.7}_{2.9}$	$76.7_{0.4}$	$\mathbf{78.8}_{0.8}$	$68.7_{0.6}$	$76.4_{0.7}$	$60.0_{1.4}$	$69.2_{0.5}$
RSQ	$\mathbf{8.051}_{.01}$	$\mathbf{68.7}_{0.2}$	$\mathbf{63.5}_{0.5}$	$\mathbf{68.5}_{0.6}$	$53.7_{1.7}$	$78.2_{2.0}$	$\mathbf{77.1}_{0.3}$	$\mathbf{78.8}_{1.0}$	$\mathbf{69.0}_{0.4}$	$\mathbf{77.1}_{0.8}$	$\mathbf{61.2}_{0.9}$	$\mathbf{69.6}_{0.3}$

Results

Table 2. Comparison of recent layer-wise quantization approaches with RSQ on multiple downstream tasks. We report perplexity for WikiText and accuracy for all other tasks. The model is quantized to 3-bit. The best-performing method across all quantization approaches is highlighted in bold. We denote the standard deviation across three runs as a subscript.

Method	Wiki	LAMBADA_{openai}	LAMBADA_{std}	WinoGrande	ArcC	ArcE	HellaSwag	PIQA	MMLU	GSM8k	TruthfulQA	Avg
LLaMA3-8B-Instruct												
Full Model	8.311	71.9	65.0	71.7	56.7	79.6	75.8	78.5	65.6	79.9	51.7	69.7
GPTQ	$10.682_{.04}$	$50.7_{2.0}$	$44.6_{1.0}$	$65.7_{0.7}$	$35.8_{1.4}$	$56.9_{2.3}$	$65.8_{0.2}$	$70.4_{1.7}$	$50.8_{0.4}$	$27.5_{1.5}$	$44.8_{1.0}$	$51.3_{0.7}$
QuaRot	$9.517_{.11}$	$68.8_{1.2}$	$59.7_{1.0}$	$70.2_{1.2}$	$49.6_{1.5}$	$74.9_{0.9}$	$71.3_{0.0}$	$76.8_{0.7}$	$57.6_{0.5}$	$61.2_{1.7}$	$48.0_{0.9}$	$63.8_{0.5}$
RSQ	$\mathbf{9.046}_{.01}$	$\mathbf{70.8}_{0.3}$	$\mathbf{62.3}_{0.4}$	$\mathbf{70.6}_{0.7}$	$\mathbf{50.3}_{1.1}$	$\mathbf{76.5}_{1.3}$	$\mathbf{72.1}_{0.2}$	$\mathbf{77.1}_{0.5}$	$\mathbf{60.0}_{0.3}$	$\mathbf{63.4}_{2.4}$	$\mathbf{50.6}_{1.7}$	$\mathbf{65.4}_{0.1}$
Mistral-NeMo-12B												
Full Model	6.095	75.8	68.3	75.1	59.1	80.1	82.2	82.1	68.2	80.8	54.8	72.7
GPTQ	$9.537_{.01}$	$44.6_{0.6}$	$29.8_{4.3}$	$57.3_{1.1}$	$38.8_{0.2}$	$58.0_{1.6}$	$60.5_{2.0}$	$71.2_{0.6}$	$48.0_{1.1}$	$37.6_{1.5}$	$47.0_{1.5}$	$49.3_{0.4}$
QuaRot	$6.782_{.00}$	$\mathbf{75.6}_{0.3}$	$60.9_{5.8}$	$72.8_{0.7}$	$\mathbf{55.8}_{1.5}$	$76.9_{1.1}$	$78.1_{0.3}$	$79.7_{0.4}$	$63.5_{0.0}$	$\mathbf{71.1}_{1.0}$	$52.5_{0.9}$	$68.7_{1.1}$
RSQ	$\mathbf{6.673}_{.01}$	$75.4_{0.3}$	$\mathbf{66.5}_{0.6}$	$\mathbf{73.5}_{0.4}$	$55.7_{0.5}$	$\mathbf{77.2}_{1.1}$	$\mathbf{78.5}_{0.1}$	$\mathbf{80.7}_{0.2}$	$\mathbf{64.3}_{0.3}$	$\mathbf{71.1}_{0.4}$	$\mathbf{52.9}_{0.5}$	$\mathbf{69.6}_{0.2}$
Qwen-2.5-7B-Instruct												
Full Model	5.335	75.2	68.6	72.7	59.0	76.4	85.2	81.0	83.3	84.4	65.5	75.1
GPTQ	$9.577_{.03}$	$53.6_{1.0}$	$48.6_{1.8}$	$62.1_{0.9}$	$47.1_{1.8}$	$66.9_{2.1}$	$72.8_{0.0}$	$73.7_{0.5}$	$60.2_{0.5}$	$40.9_{10.5}$	$53.7_{1.5}$	$58.0_{1.8}$
QuaRot	$8.053_{.02}$	$67.9_{0.3}$	$62.1_{0.4}$	$68.3_{1.2}$	$\mathbf{54.0}_{1.6}$	$\mathbf{78.7}_{2.9}$	$76.7_{0.4}$	$\mathbf{78.8}_{0.8}$	$68.7_{0.6}$	$76.4_{0.7}$	$60.0_{1.4}$	$69.2_{0.5}$
RSQ	$\mathbf{8.051}_{.01}$	$\mathbf{68.7}_{0.2}$	$\mathbf{63.5}_{0.5}$	$\mathbf{68.5}_{0.6}$	$53.7_{1.7}$	$78.2_{2.0}$	$\mathbf{77.1}_{0.3}$	$\mathbf{78.8}_{1.0}$	$\mathbf{69.0}_{0.4}$	$\mathbf{77.1}_{0.8}$	$\mathbf{61.2}_{0.9}$	$\mathbf{69.6}_{0.3}$

Results on long-context tasks

Table 3. Comparison of RSQ and QuaRot across multiple long-context benchmarks using three different calibration dataset configurations. The model is quantized to 3-bit. The best-performing method among all quantization approaches is highlighted in bold.

Method	LITM				L-Eval								LongICLBench			LongCodeArena
	P=1	*P*=15	*P*=30	Avg	TOEFL	QuALITY	Coursera	SFiction	GSM	CodeU	TopicRet	Avg	Banking77	TecRED	Avg	CodeGen
Full Model	53.67	45.61	46.33	48.54	81.04	60.40	52.62	71.88	81.00	4.44	64.67	59.43	59.40	41.65	50.52	0.298
number of samples = 256, sequence length = 4096																
QuaRot	$50.80_{1.6}$	$43.47_{2.4}$	$43.29_{2.5}$	$45.85_{2.0}$	$72.24_{0.9}$	$52.15_{0.8}$	$46.90_{0.7}$	$\mathbf{65.62}_{3.1}$	$63.33_{6.6}$	$2.22_{0.9}$	$45.11_{3.5}$	$49.65_{1.4}$	$\mathbf{62.40}_{4.3}$	$44.76_{5.2}$	$53.58_{3.2}$	$0.206_{.013}$
RSQ	$\mathbf{52.23}_{0.5}$	$\mathbf{46.03}_{2.7}$	$\mathbf{45.59}_{3.5}$	$\mathbf{47.95}_{2.0}$	$\mathbf{76.21}_{2.1}$	$\mathbf{54.46}_{0.7}$	$\mathbf{49.27}_{1.3}$	$63.80_{4.0}$	$\mathbf{66.33}_{3.8}$	$\mathbf{4.07}_{0.5}$	$\mathbf{50.89}_{2.4}$	$\mathbf{52.14}_{0.3}$	$58.60_{7.7}$	$\mathbf{49.83}_{0.7}$	$\mathbf{54.21}_{3.5}$	$\mathbf{0.231}_{.008}$
number of samples = 512, sequence length = 2048																
QuaRot	$49.17_{1.2}$	$44.32_{1.0}$	$43.51_{0.5}$	$45.67_{0.2}$	$71.13_{2.3}$	$51.16_{1.4}$	$46.95_{0.7}$	$58.33_{0.9}$	$65.67_{4.9}$	$\mathbf{4.81}_{2.1}$	$48.00_{7.0}$	$49.43_{2.0}$	$65.40_{0.6}$	$45.48_{3.1}$	$\mathbf{55.44}_{1.8}$	$0.205_{.007}$
RSQ	$\mathbf{51.39}_{1.6}$	$\mathbf{45.89}_{0.8}$	$\mathbf{45.87}_{1.2}$	$\mathbf{47.72}_{0.8}$	$\mathbf{73.23}_{2.2}$	$\mathbf{53.96}_{1.7}$	$\mathbf{48.79}_{1.0}$	$\mathbf{61.98}_{2.6}$	$\mathbf{70.00}_{2.1}$	$4.44_{0.0}$	$\mathbf{51.78}_{7.0}$	$\mathbf{52.02}_{1.9}$	$62.33_{2.2}$	$\mathbf{46.17}_{5.5}$	$54.25_{3.8}$	$\mathbf{0.227}_{.002}$
number of samples = 1024, sequence length = 1024																
QuaRot	$48.66_{2.6}$	$47.20_{0.9}$	$\mathbf{48.80}_{1.7}$	$48.22_{1.4}$	$73.11_{2.1}$	$51.82_{1.4}$	$48.64_{1.8}$	$62.24_{1.3}$	$65.00_{0.8}$	$2.59_{1.3}$	$40.45_{3.0}$	$49.12_{0.9}$	$59.73_{8.0}$	$46.02_{7.0}$	$52.87_{6.9}$	$0.220_{.001}$
RSQ	$\mathbf{51.10}_{0.2}$	$\mathbf{48.01}_{0.7}$	$48.62_{1.3}$	$\mathbf{49.24}_{0.2}$	$\mathbf{74.72}_{2.8}$	$\mathbf{55.78}_{2.0}$	$\mathbf{50.58}_{0.8}$	$\mathbf{64.85}_{3.3}$	$\mathbf{69.33}_{1.2}$	$\mathbf{3.70}_{0.5}$	$\mathbf{47.55}_{7.4}$	$\mathbf{52.35}_{2.0}$	$\mathbf{61.20}_{5.2}$	$\mathbf{46.30}_{1.9}$	$\mathbf{53.75}_{3.5}$	$\mathbf{0.225}_{.002}$

RSQ is a plug-and-play module

Table 6: Joint quantization.

Method	Tasks	
	Wiki PPL ↓	Avg Acc (%) ↑
W4A4KV16		
QuaRot	$10.26_{.05}$	$63.7_{0.1}$
RSQ	$\mathbf{10.03}_{.04}$	$\mathbf{64.3}_{0.1}$
W4A4KV4		
QuaRot	$12.72_{.16}$	$57.6_{0.7}$
RSQ	$\mathbf{12.32}_{.20}$	$\mathbf{58.3}_{0.6}$

RSQ is a plug-and-play module

Table 6: Joint quantization.

Method	Tasks	
	Wiki PPL ↓	Avg Acc (%) ↑
W4A4KV16		
QuaRot	$10.26_{.05}$	$63.7_{0.1}$
RSQ	$\mathbf{10.03}_{.04}$	$\mathbf{64.3}_{0.1}$
W4A4KV4		
QuaRot	$12.72_{.16}$	$57.6_{0.7}$
RSQ	$\mathbf{12.32}_{.20}$	$\mathbf{58.3}_{0.6}$

Table 7: RSQ + SOTA 2-bit VQ.

Method	Tasks		
	C4 PPL ↓	Wiki PPL ↓	Avg Acc (%) ↑
QuIP#	11.62	9.029	54.92
RSQ (w/ QuIP#)	**8.622**	**6.589**	**60.69**
QTIP	9.037	6.864	62.38
RSQ (w/ QTIP)	**8.167**	**6.244**	**65.70**

RSQ on 70B model

Table 5: Comparison RSQ against QuaRot on multiple downstream tasks using LLaMA3-70B-Instruct. Results are run with seed 0.

Method	Wiki	LAMBADA$_{oai}$	LAMBADA$_{std}$	WinoGrande	ArcC	ArcE	HellaSwag	PIQA	MMLU	GSM8k	TruthfulQA	Avg
Full Model	5.399	76.1	72.4	76.2	64.5	84.6	82.5	81.9	79.8	92.1	61.6	77.17
					4-bit							
QuaRot	5.882	76.0	71.9	75.4	64.1	84.5	81.7	82.0	77.8	91.1	61.8	76.63
RSQ	**5.730**	76.0	72.4	75.8	64.2	84.3	81.9	81.7	79.0	91.5	60.5	**76.73**
					3-bit							
QuaRot	7.511	74.2	69.0	72.3	59.2	80.9	78.4	79.2	70.4	85.1	57.2	72.59
RSQ	**6.503**	76.8	71.7	75.5	62.1	82.9	80.9	81.0	76.0	89.8	60.8	**75.75**
					2-bit							
QuaRot	26.17	34.7	24.7	52.3	27.7	49.2	41.6	62.5	26.5	2.1	40.4	36.17
RSQ	**10.623**	70.2	60.2	63.5	51.0	76.5	69.8	75.7	56.7	48.2	46.6	**61.84**

Summary

- Improved layer-wise pruning and quantization by “Token importance”
- These methods drastically reduce the model size
  - For example, 4-bit model reduces the 16-bit model size (and speed-up) by **4x**, with a minimal performance drop!

Summary

- Improved layer-wise pruning and quantization by “Token importance”
- These methods drastically reduce the model size
  - For example, 4-bit model reduces the 16-bit model size (and speed-up) by **4x**, with a minimal performance drop!

Questions?

Thank you!

Feel free to reach out to ylsung@cs.unc.edu if you have questions!

My website: https://ylsung.github.io/

Support Slides

Theoretical Results

$$||\mathcal{L}_g(\hat{\mathbf{W}}^{(l)}) - \mathcal{L}_g(\mathbf{W}^{(l)})||_2^2 \approx ||\frac{\partial \mathcal{L}}{\partial \mathbf{X}^{(l+1)}} \cdot (\hat{\mathbf{W}}^{(l)}\mathbf{X}^{(l)} - \mathbf{W}^{(l)}\mathbf{X}^{(l)})||_2^2 \tag{3}$$

$$= \sum_{i=0}^{T}\sum_{j=0}^{dout} ||\frac{\partial \mathcal{L}}{\partial \mathbf{X}^{(l+1)}}|_{j,i} \times (\hat{\mathbf{W}}^{(l)}\mathbf{X}^{(l)} - \mathbf{W}^{(l)}\mathbf{X}^{(l)})_{j,i}||_2^2 \tag{4}$$

Theoretical Results

$$||\mathcal{L}_g(\hat{\mathbf{W}}^{(l)}) - \mathcal{L}_g(\mathbf{W}^{(l)})||_2^2 \approx ||\frac{\partial \mathcal{L}}{\partial \mathbf{X}^{(l+1)}} \cdot (\hat{\mathbf{W}}^{(l)}\mathbf{X}^{(l)} - \mathbf{W}^{(l)}\mathbf{X}^{(l)})||_2^2 \tag{3}$$

$$= \sum_{i=0}^{T}\sum_{j=0}^{dout} ||\frac{\partial \mathcal{L}}{\partial \mathbf{X}^{(l+1)}}|_{j,i} \times (\hat{\mathbf{W}}^{(l)}\mathbf{X}^{(l)} - \mathbf{W}^{(l)}\mathbf{X}^{(l)})_{j,i}||_2^2 \tag{4}$$

$$\sum_{i=0}^{T}\sum_{j=0}^{dout} ||\frac{\partial \mathcal{L}}{\partial \mathbf{X}^{(l+1)}}|_{j,i} \times (\hat{\mathbf{W}}^{(l)}\mathbf{X}^{(l)} - \mathbf{W}^{(l)}\mathbf{X}^{(l)})_{j,i}||_2^2 \tag{5}$$

$$\approx \sum_{i=0}^{T}\sum_{j=0}^{dout} || \mathbf{r}_{\mathbf{token}i} \times \mathbf{r}_{\mathbf{channel}j} \times (\hat{\mathbf{W}}^{(l)}\mathbf{X}^{(l)} - \mathbf{W}^{(l)}\mathbf{X}^{(l)})_{j,i}||_2^2 \tag{6}$$

$$\overset{\text{if GPTQ}}{=} \sum_{i=0}^{T}\sum_{j=0}^{dout} || \mathbf{r}_{\mathbf{token}i} \times (\hat{\mathbf{W}}^{(l)}\mathbf{X}^{(l)} - \mathbf{W}^{(l)}\mathbf{X}^{(l)})_{j,i}||_2^2 \tag{7}$$

Ablation on importance measure

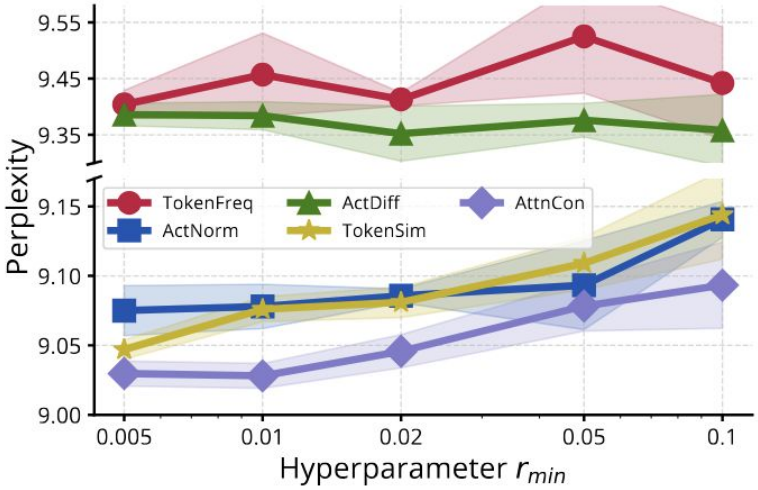


Figure 1: Evaluation of five dynamic approaches with varying r_{min}.

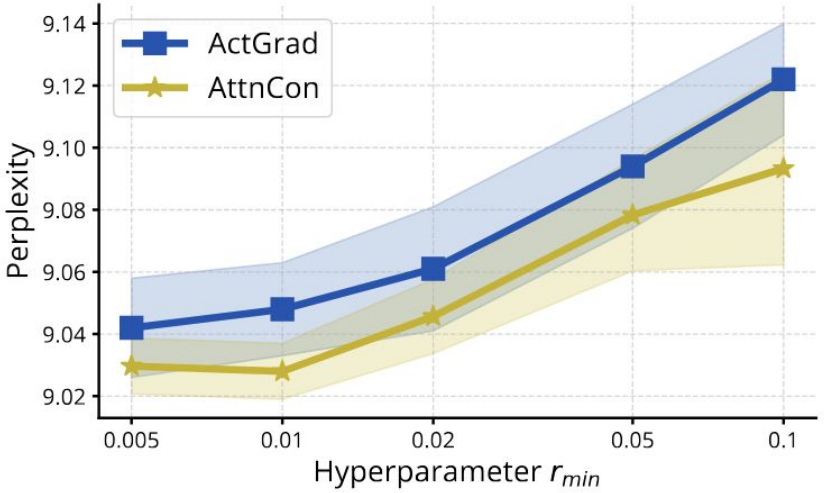


Figure 2: Comparison between AttnCon and ActGrad (reach 9.036 using original values).