

Published as a conference paper at ICLR 2025

SPINQUANT: LLM QUANTIZATION WITH LEARNED ROTATIONS

Zechun Liu* **Changsheng Zhao*** **Igor Fedorov** **Bilge Soran** **Dhruv Choudhary**
Raghuraman Krishnamoorthi **Vikas Chandra** **Yuandong Tian** **Tijmen Blankevoort**
Meta

Background

- 大语言模型 (LLM) 参数量巨大 (如 LLaMA-2 70B 需数百 GB 显存)，推理成本极高；限制了在资源受限设备上的部署；
- 量化技术将模型权重与激活从高精度 (FP16/FP32) 压缩到低比特 (如 INT8、INT4) 以节省资源；
 - 显著降低显存占用 (FP16 $\rightarrow$ INT4 节省 4 $\times$)
 - 加速推理速度 (可在整数矩阵乘法上执行)
 - 降低功耗、支持端侧部署

Existing situation

方法	核心思路	局限性
SmoothQuant (ICLR 2023)	在权重和激活间“平衡缩放” outlier (通过比例转移)	无法彻底消除 outlier, 仅改善部分层; 需调整尺度参数
LLM-QAT (Meta 2023)	模拟量化噪声进行量化感知训练 (QAT)	训练成本高, 需反向传播; 不适合无数据 PTQ 场景
QuIP# / QuaRot (2023–2024)	使用随机旋转矩阵 (Hadamard/正交) 打散 outlier 分布	随机性强 → 量化精度方差大 (可达 13pt)
AWQ / OmniQuant	权重重标定或混合精度策略	主要针对权重量化, 对激活和 KV-cache 效果有限

Challenge

- 量化的核心挑战：Outliers（异常值）
- 当存在 outlier 时，量化范围（min, max）被极端值“拉宽”；
- 结果：大部分普通值都被映射到量化区间的狭小区域；
- 导致有效位数减少、信息丢失严重 → 精度骤降。

$$\alpha = \frac{X_{max} - X_{min}}{q_{max} - q_{min}}$$

$$X_Q = \alpha \left\lfloor \frac{X_R - \beta}{\alpha} \right\rfloor + \beta$$

3bit: $x_1 \quad x_2 \quad x_3 \quad x_4 \quad x_5$
 $[0.3, 0.5, 0.4, 0.1, 0.8]$ 无异常值

$$a = \left(\frac{0.8 - 0.1}{2^3 - 1} \right) = \frac{0.7}{7} = 0.1$$

$$q_{x_1} = \text{round} \left(\frac{x_1 - 0.1}{a} \right) \approx \frac{0.3 - 0.1}{0.1} = 2$$

$$\text{反量化: } q_{x_1} \cdot a + 0.1 = 0.3$$

$x_1 \quad x_2 \quad x_3 \quad x_4 \quad x_5 \quad x_6$
 $[0.4, 0.2, 0.8, 4, 0.6, 21.2]$ 有异常值

$$a = \frac{21.2 - 0.2}{2^3 - 1} = \frac{21}{7} = 3$$

$$q_{x_1} = \text{round} \left(\frac{0.4 - 0.2}{a} \right) \approx \frac{0.2}{3} = 0$$

$$\text{反量化: } q_{x_1} \cdot a + 0.2 = \textcircled{0.2}$$

$$q_{x_4} = \text{round} \left(\frac{4 - 0.2}{a} \right) = \frac{3.8}{3} = 3$$

$$\text{反量化: } q_{x_4} \cdot a + 0.2 = \textcircled{3.2}$$

$$\alpha = \frac{X_{max} - X_{min}}{q_{max} - q_{min}}$$

$$X_Q = \alpha \left[\frac{X_R - \beta}{\alpha} \right] + \beta$$

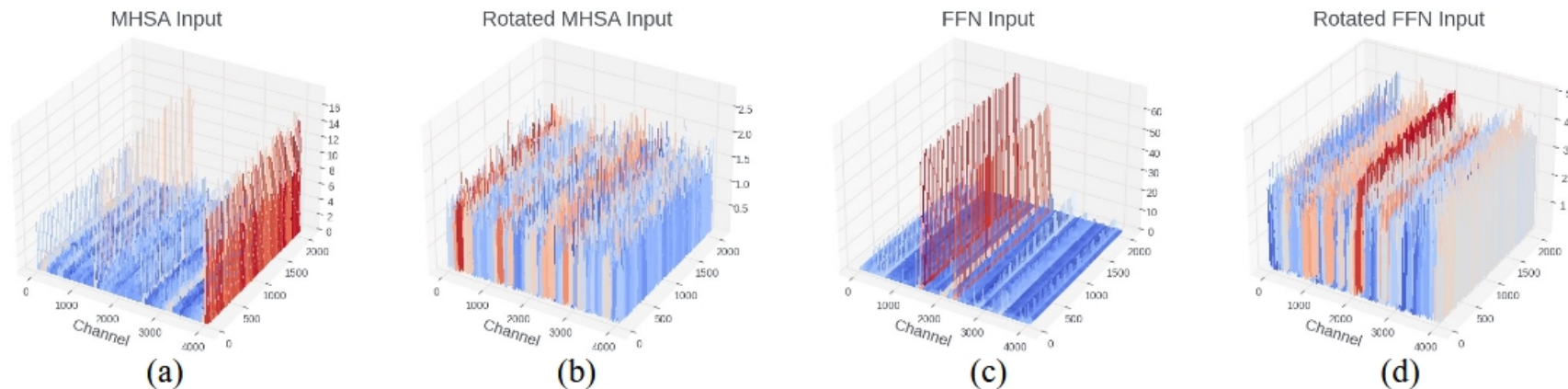


Figure 2: Activation distribution in LLaMA-2 7B model before and after rotation. Outliers exist in particular channels before rotation. Since channel-wise quantization is not supported in most hardware, outlier removal using rotation enables accurate token-wise or tensor-wise quantization.

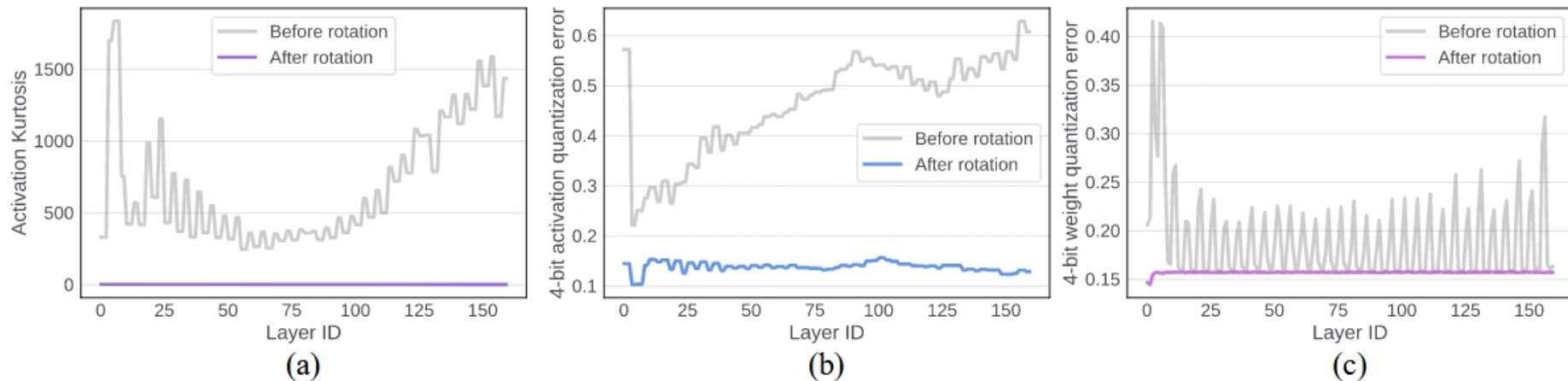


Figure 3: Outlier measurement and quantization error across input activation and weights in the five layers that take inputs from the residual (Q/K/V/Up/Gate-projection) of each block in the LLaMA-2 7B model. (a) After rotation, *kurtosis* of activation distributions is significantly reduced to approximately three across all layers. Quantization error is reduced after rotation in both (b) activations and (c) weights.

SpinQuant 的切入点

论文从 “旋转不变性 (rotational invariance)” 出发提出新思路：

- Transformer 的残差流允许引入旋转矩阵R，且在浮点数精度下网络输出保持完全一致；
- 通过对激活或权重施加旋转，可以重新分布值域，让异常值与普通值 “混合” ；
- 从而使得激活分布更加平滑、易于量化。
- 将这些旋转矩阵作为可训练参数，通过优化让量化误差最小化

$$\min_{R_1, R_2} \|Q(X; R_1, R_2) - F(X)\|^2$$

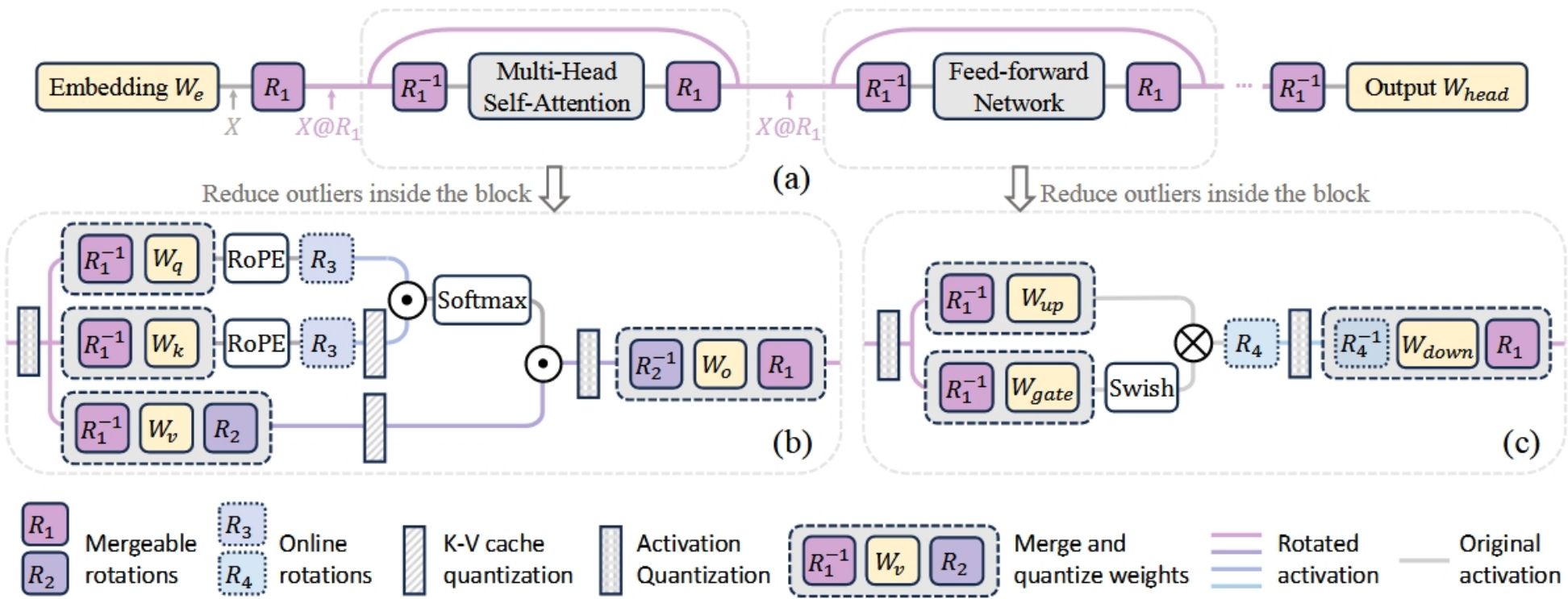


Figure 1: Overall diagram of rotation. (a) The residual stream can be rotated in the transformer network, resulting in numerically equivalent floating point networks before and after rotation. The rotated activations exhibit fewer outliers and are easier to quantize. (b) & (c) The rotation matrix can be integrated with the corresponding weight matrices and we further define R_2 , R_3 , and R_4 for reducing outliers inside the block.

Experiment

Table 1: Comparison of the perplexity score on WikiText2 and averaged accuracy on eight Zero-shot Common Sense Reasoning tasks. Results for SmoothQuant (Xiao et al., 2022), LLM-QAT (Liu et al., 2023c), GPTQ (Frantar et al., 2022) were obtained using their publicly released codebase. While OmniQuant (Shao et al., 2023), AWQ (Lin et al., 2023), and QuIP# (Tseng et al., 2024) results were quoted from their papers. Full results are in the Appendix.

#Bits (W-A-KV)	Method	LLaMA-2 7B		LLaMA-2 13B		LLaMA-2 70B		LLaMA-3.2 1B		LLaMA-3.2 3B		LLaMA-3 8B		Mistral-7B	
		0-shot ⁸	Wiki	0-shot ⁸	Wiki	0-shot ⁸	Wiki	0-shot ⁸	Wiki	0-shot ⁸	Wiki	0-shot ⁸	Wiki	0-shot ⁸	Wiki
		Avg.(↑)	(↓)	Avg.(↑)	(↓)	Avg.(↑)	(↓)	Avg.(↑)	(↓)	Avg.(↑)	(↓)	Avg.(↑)	(↓)	Avg.(↑)	(↓)
16-16-16	FloatingPoint	66.9	5.5	68.3	5.0	72.9	3.3	56.9	13.4	63.9	10.7	69.6	6.1	71.0	5.4
	RTN	62.4	7.9	57.3	6.7	68.6	5.0	55.4	20.7	58.6	29.0	65.5	8.2	59.3	6.8
	SmoothQuant	58.9	7.5	63.6	6.1	70.6	4.1	47.1	1e2	55.6	3e2	61.0	10.7	—	—
	LLM-QAT	64.8	11.4	67.5	14.5	—	—	53.2	21.0	60.8	41.1	67.2	7.7	—	—
	AWQ (w4)	—	6.2	—	5.1	—	—	—	—	—	—	—	—	—	—
	OmniQuant (w4)	—	5.7	—	5.0	—	3.5	—	—	—	—	—	—	—	—
	QuIP# (w4)	—	5.6	—	5.0	—	3.4	—	—	—	—	—	—	—	—
	GPTQ	64.9	20.2	65.2	5.9	71.7	4.3	55.0	17.3	58.7	25.2	64.5	7.2	51.7	8.6
4-8-16	SpinQuant _{no had}	65.7	5.8	68.2	5.1	72.1	3.7	56.0	15.3	61.4	11.6	68.6	6.7	68.8	5.7
	SpinQuant _{had}	65.7	5.7	68.1	5.0	72.7	3.5	56.5	14.4	63.2	11.5	68.4	6.5	69.9	5.5
	RTN	62.5	7.9	57.6	6.7	68.4	5.0	55.7	20.7	58.4	28.8	65.3	8.2	58.9	6.7
	SmoothQuant	58.8	7.5	63.4	6.1	70.5	4.1	47.1	1e2	55.5	3e2	60.9	10.7	—	—
	LLM-QAT	64.6	11.4	67.5	14.2	—	—	53.1	21.0	60.5	39.3	66.9	7.6	—	—
	GPTQ	64.8	20.2	65.3	5.9	71.6	4.3	54.8	17.3	58.7	24.1	64.6	7.2	51.7	8.6
	SpinQuant _{no had}	65.8	5.8	68.1	5.1	72.2	3.7	55.7	15.3	61.8	11.7	68.6	6.7	69.4	5.7
	SpinQuant _{had}	65.8	5.7	68.2	5.1	72.7	3.5	55.8	14.3	63.2	11.2	68.8	6.5	70.2	5.5
4-8-8	RTN	35.6	2e3	35.3	7e3	35.1	2e5	41.2	1e2	42.1	7e2	43.9	2e2	41.4	4e2
	SmoothQuant	41.8	3e2	44.9	34.5	57.7	57.1	37.9	2e3	43.6	4e2	40.3	9e2	—	—
	LLM-QAT	47.8	12.9	34.3	4e3	—	—	42.0	62.1	46.9	37.6	44.9	42.9	—	—
	GPTQ	36.8	9e3	35.2	5e3	35.5	2e6	41.6	1e2	43.4	3e2	40.6	2e2	40.4	3e2
	SpinQuant _{no had}	57.0	9.2	61.8	7.2	61.0	7.3	44.8	48.4	52.9	22.4	51.9	18.6	52.7	13.4
	SpinQuant _{had}	64.1	5.9	67.2	5.2	71.0	3.8	53.5	15.3	61.0	11.1	65.8	7.1	68.4	5.7
	RTN	37.1	2e3	35.5	7e3	35.0	2e5	40.6	2e2	41.2	8e2	43.1	3e2	41.4	4e2
	SmoothQuant	39.0	7e2	40.5	56.6	55.9	10.5	36.5	2e3	40.0	6e2	38.7	2e3	—	—
4-4-16	LLM-QAT	44.9	14.9	35.0	4e3	—	—	41.5	76.2	45.9	42.0	43.2	52.5	—	—
	GPTQ	36.8	9e3	35.2	5e3	35.6	1e6	41.6	1e2	41.1	4e2	40.5	2e2	41.3	2e2
	SpinQuant _{no had}	56.0	9.2	60.7	7.1	62.0	7.4	45.3	47.7	52.9	22.4	52.6	18.6	52.4	13.7
	SpinQuant _{had}	64.0	5.9	66.9	5.3	71.2	3.8	53.4	15.9	60.5	11.4	65.5	7.3	68.6	5.8

Experiment

Table 2: Compared to Hadamard rotation, SpinQuant learned rotation consistently outperform by a significant margin. Results are averaged accuracy on eight Zero-shot CommonSense Reasoning tasks.

	LLaMA-3.2 3B		LLaMA-3 8B		Mistral-7B	
	4-4-16	4-4-4	4-4-16	4-4-4	4-4-16	4-4-4
Random Hadamard $R_{\{1,2\}}$	49.8	49.6	49.5	50.0	51.4	51.5
SpinQuant _{no had} $R_{\{1,2\}}$	52.9(↑3.1)	52.9(↑3.3)	51.9(↑2.4)	52.6(↑2.5)	52.7(↑1.3)	52.4(↑0.9)
Random Hadamard $R_{\{1,2,3,4\}}$	59.0	58.4	64.2	63.9	52.7	52.4
SpinQuant _{had} $R_{\{1,2,3,4\}}$	61.0(↑2.1)	60.5(↑2.2)	65.8(↑1.6)	65.5(↑1.6)	68.4(↑15.7)	68.6(↑16.2)

Ablation Study

Table 5: Comparison with QuaRot (Ashkboos et al., 2023b).

	LLaMA-3 8B (FP: 69.6, 6.1)				LLaMA-3 70B (FP: 74.5, 2.8)			
	4-4-16		:	4-4-4	4-4-16		:	4-4-4
	0-shot ^s Avg.(↑)	Wiki (↓)	:	0-shot ^s Avg.(↑)	Wiki (↓)	:	0-shot ^s Avg.(↑)	Wiki (↓)
QuaRot+RTN	59.5	10.4	:	58.6	10.9	:	41.5	91.2
SpinQuant _{had} +RTN	64.6	7.7	:	64.1	7.8	:	70.1	4.1
QuaRot+GPTQ	63.8	7.9	:	63.3	8.0	:	65.4	20.4
SpinQuant _{had} +GPTQ	65.8	7.1	:	65.5	7.3	:	69.5	5.5