SpinQuant

SpinQuant- 用旋转矩阵去除LLM outliers的量化方法，其上限明显高于GPTQ和SmoothQuant两类方法

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

Abstract

Post-training quantization (PTQ) techniques applied to weights, activations, and the KV cache greatly reduce memory usage, latency, and power consumption of Large Language Models (LLMs), but may lead to large quantization errors when outliers are present. Recent findings suggest that rotating activation or weight matrices helps remove outliers and benefits quantization. In this work, we identify a collection of applicable rotation parameterizations that lead to identical outputs in full-precision Transformer architectures, and find that some random rotations lead to much better quantization than others, with an up to *13 points* difference in downstream zero-shot reasoning performance. As a result, we propose SpinQuant that *optimizes* (or *learns*) the rotation matrices with *Cayley* optimization on a small validation set. With 4-bit quantization of weight, activation, and KV-cache, SpinQuant narrows the accuracy gap on zero-shot reasoning tasks with full precision to merely 2.9 points on the LLaMA-2 7B model, surpassing LLM-QAT by 19.1 points and SmoothQuant by 25.0 points. SpinQuant also outperforms concurrent work QuaRot, which applies random rotations to remove outliers. In particular, for LLaMA-2 7B/LLaMA-3 8B models that are hard to quantize, SpinQuant reduces the gap to full precision by 30.2%/34.1% relative to QuaRot.

- 链接：https://arxiv.org/abs/2405.16406
- 单位：Meta

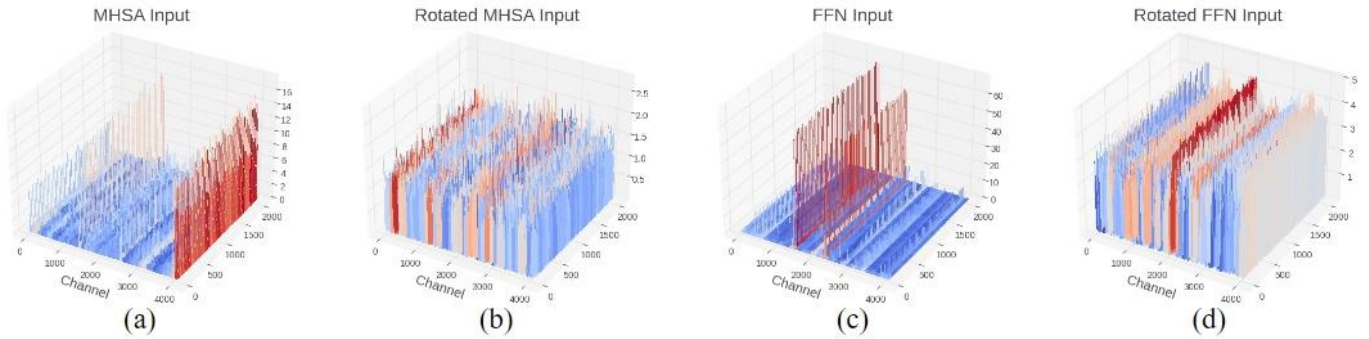
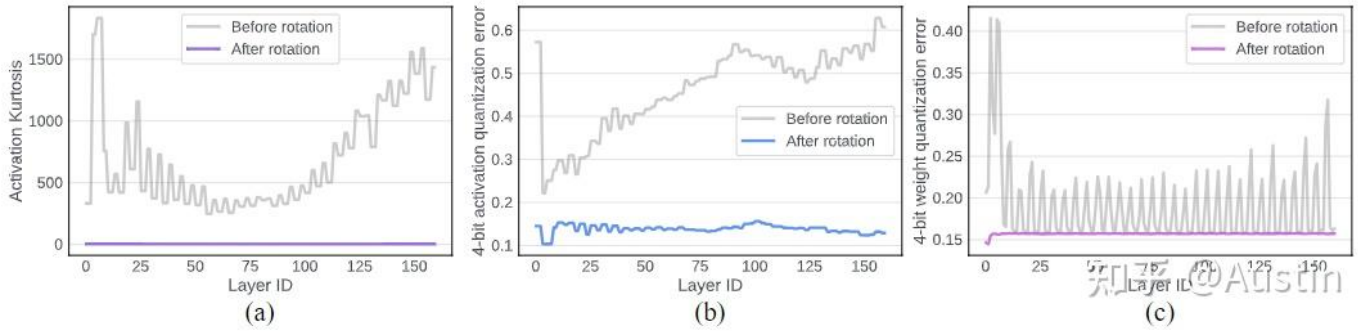


Figure 2: Activation distribution in LLaMA-2 7B model before and after rotation. Outliers exist in particular channels before rotation. Since channel-wise quantization is not supported in most hardware, outlier removal using rotation enables accurate token-wise or tensor-wise quantization.



SpinQuant用旋转矩阵消除outliers

旋转矩阵的优点有很多，例如正交性（不会改变模的大小，不会影响RMSNorm分母）、逆等于其转置等。

ROPE旋转位置编码已经证明了自己。

SpinQuant使用Rotation去smooth LLMs中的离群点真不错，上限明显高于GPTQ和Smoothquan这两类方法，且Smooth方法是其子集。

这篇文章理解起来还是比较容易的，下面一张图即可搞定：

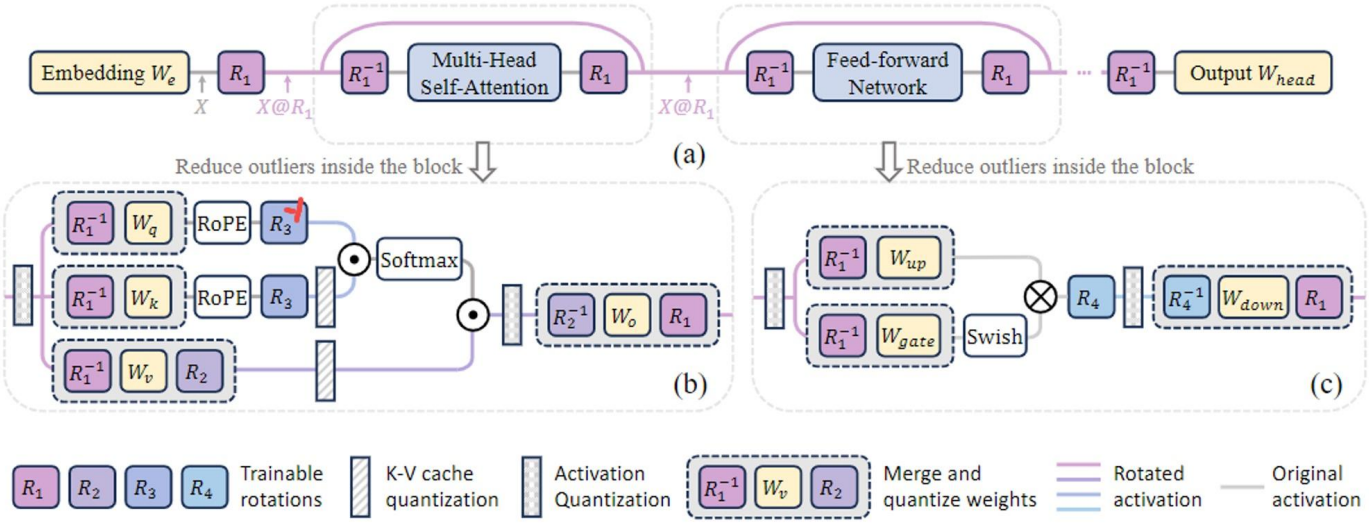
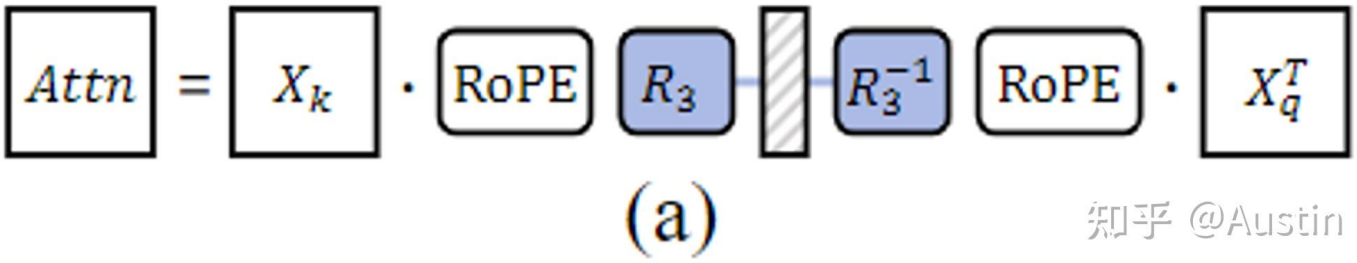


Figure 1: **Overall diagram of rotation.** (a) The residual stream can be rotated in the transformer network, resulting in numerically equivalent floating point networks before and after rotation. The rotated activations exhibit fewer outliers and are easier to quantize. (b) & (c) The rotation matrix 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.

Figure 1中几点说明：

- Query的R_3画错了，应该是 R_3^{-1}，否则相乘后不等价，如下图所示。



- R_4由于前面是Element-wise相乘，因此不能融合到 W_{up}；如果把ROPE扩展为稀疏矩阵，是不是就可以和 R_3融合了？（在最大长度不超过ROPE的初始值下）

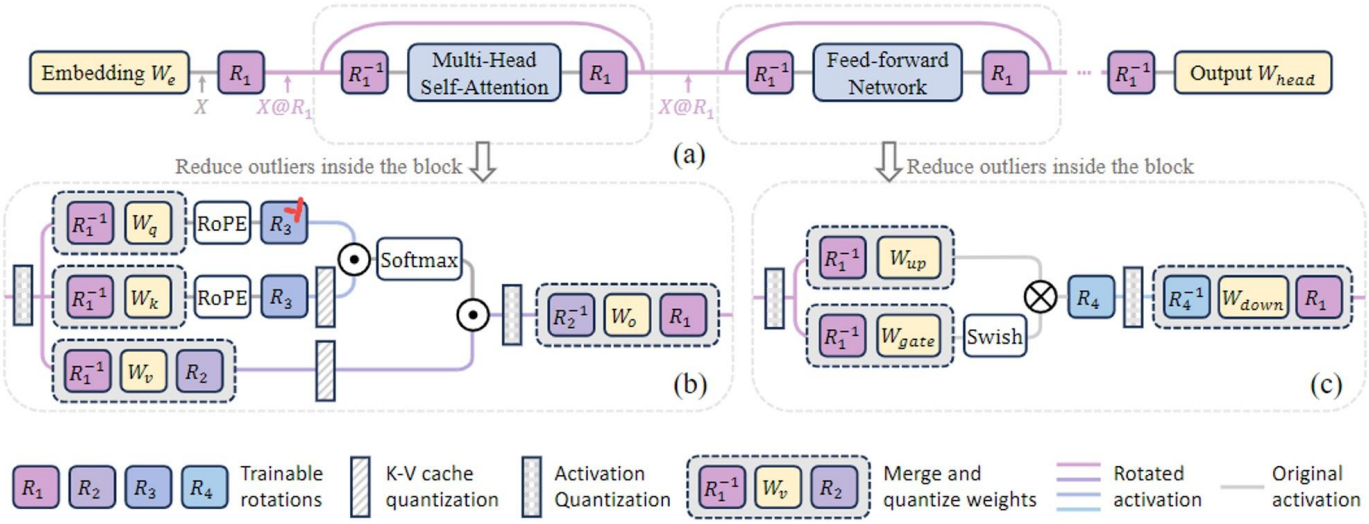


Figure 1: **Overall diagram of rotation.** (a) The residual stream can be rotated in the transformer network, resulting in numerically equivalent floating point networks before and after rotation. The rotated activations exhibit fewer outliers and are easier to quantize. (b) & (c) The rotation matrix 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.

- 虽然旋转矩阵是学习出来的（如下图），但相比GPTQ这类直接重建weight，可能会改变模型的泛化性（例如过拟合），但是SpinQuant不会改变模型参数，只影响量化性能，因此上限高于GPTQ和Smoothquant。

$$\arg\min_{R \in \mathcal{M}} \mathcal{L}_Q(R_1, R_2 \mid W, X)$$

We apply the *Cayley SGD* method to solve Eqn. (1) for $\{R_1, R_2\}$, while the underlying weight parameters in the network remain frozen. $\{R_1, R_2\}$ count for only 0.26% of the weight size and is constrained to be orthonormal. Consequently, the underlying floating-point network remains unchanged, and the rotation only influences the quantization performance.

SpinQuant的效果还不错，且更新了LLaMA-3的效果。

#Bits W-A-KV	Method	LLaMA-3 8B 0-shot[8] Avg.(↑)	LLaMA-3 8B Wiki (↓)	LLaMA-3 70B 0-shot[8] Avg.(↑)	LLaMA-3 70B Wiki (↓)	LLaMA-2 7B 0-shot[4] Avg.(↑)	LLaMA-2 7B 0-shot[8] Avg.(↑)	LLaMA-2 7B Wiki (↓)	LLaMA-2 13B 0-shot[4] Avg.(↑)	LLaMA-2 13B 0-shot[8] Avg.(↑)	LLaMA-2 13B Wiki (↓)	LLaMA-2 70B 0-shot[4] Avg.(↑)	LLaMA-2 70B 0-shot[8] Avg.(↑)	LLaMA-2 70B Wiki (↓)
16-16-16	FloatingPoint	69.6	6.1	74.5	2.8	68.6	66.9	5.5	69.9	68.3	5.0	76.0	72.9	3.3
4-16-16	RTN	65.4	7.8	35.5	1e5	65.3	63.6	7.2	59.9	57.9	6.4	72.3	69.2	4.6
	SmoothQuant	61.0	10.7	66.9	12.0	62.3	59.1	7.5	66.1	63.3	6.1	73.6	70.2	4.1
	LLM-QAT	67.7	7.1	–	–	67.1	64.9	5.9	–	–	–	–	–	–
	OmniQuant	–	–	–	–	62.5	–	5.7	64.9	–	5.0	71.1	–	3.5
	QuIP	–	–	–	–	–	–	–	66.7	–	–	69.4	–	–
	AQLM	–	–	–	–	63.6	–	–	66.3	–	–	71.9	–	–
	QuIP#	–	–	–	–	63.9	–	–	67.1	–	–	71.8	–	–
	AWQ	–	–	–	–	–	–	6.2	–	–	5.1	–	–	–
	GPTQ	66.5	7.2	35.7	1e5	66.6	64.5	11.3	67.0	64.7	5.6	75.0	71.9	3.9
	QuaRot*	66.2	7.5	57.2	41.6	64.5	62.4	6.9	69.3	67.1	5.5	74.6	71.8	3.7
	QuaRot	68.4	**6.4**	70.3	7.9	67.6	65.8	5.6	70.3	68.3	5.0	75.2	72.2	**3.5**
	SpinQuant*	67.6	6.5	71.4	**3.9**	66.7	64.6	**5.5**	69.6	67.4	**4.9**	74.9	72.2	**3.5**
	SpinQuant	**68.5**	**6.4**	**71.6**	4.8	**67.7**	**65.9**	5.6	**70.5**	**68.5**	5.0	**75.4**	**72.6**	**3.5**
4-4-16	RTN	38.5	9e2	35.6	1e5	37.3	35.6	2e3	37.9	35.3	7e3	37.2	35.1	2e5
	SmoothQuant	40.3	8e2	55.3	18.0	44.2	41.8	2e2	46.4	44.9	34.5	46.8	64.6	57.1
	LLM-QAT	44.9	42.9	–	–	48.8	47.8	12.9	–	–	–	–	–	–
	OmniQuant	–	–	–	–	–	–	14.3	–	–	12.3	–	–	–
	GPTQ	37.0	9e2	35.3	1e5	38.3	36.8	8e3	37.7	35.3	5e3	37.8	35.5	2e6
	QuaRot*	59.5	10.4	41.5	91.2	60.9	59.0	8.2	66.9	64.8	6.1	72.6	69.7	4.2
	QuaRot	63.8	7.9	65.4	20.4	65.6	63.5	6.1	68.5	66.7	5.4	72.9	70.4	3.9
	SpinQuant*	64.6	7.7	**70.1**	**4.1**	63.6	61.8	6.1	67.7	65.8	5.4	**73.5**	**71.1**	3.9
	SpinQuant	**65.8**	**7.1**	69.5	5.5	**65.9**	**64.1**	**5.9**	**69.3**	**67.2**	**5.2**	**73.5**	71.0	**3.8**
4-4-4	RTN	38.2	1e3	35.2	1e5	38.2	37.1	2e3	37.2	35.4	7e3	37.4	35.0	2e5
	SmoothQuant	38.7	1e3	52.4	22.1	40.1	39.0	6e2	42.7	40.5	56.6	58.8	55.9	10.5
	LLM-QAT	43.2	52.5	–	–	45.5	44.9	14.9	–	–	–	–	–	–
	GPTQ	37.1	1e3	35.1	1e5	38.1	36.8	9e3	37.4	35.2	5e3	37.8	35.6	1e6
	QuaRot*	58.6	10.9	41.3	92.4	60.5	58.7	8.2	66.5	64.4	6.2	72.2	69.5	4.2
	QuaRot	63.3	8.0	65.1	20.2	64.4	62.5	6.4	68.1	66.2	5.4	73.0	70.3	3.9
	SpinQuant*	64.1	7.8	**70.1**	**4.1**	63.2	61.5	6.2	67.7	65.5	5.4	73.2	70.5	3.9
	SpinQuant	**65.2**	**7.3**	69.3	5.5	**66.0**	**64.0**	**5.9**	**68.9**	**66.9**	**5.3**	**73.7**	**71.2**	**3.8**