RPTQ: Reorder-based Post-training Quantization for Large Language Models

- 2023/10/16
- Ming Wang

Overview

规格严格 功夫到家

1920 —— 2020

RPTQ: Reorder-based Post-training Quantization for Large Language Models

Zhihang Yuan* Houmo AI

Lin Niu* **Jiawei Liu** **Wenyu Liu** **Xinggang Wang†** Huazhong University of Science & Technology

Yuzhang Shang Illinois Institute of Technology

Guangyu Sun Peking University

Qiang Wu Houmo AI

Jiaxiang Wu Tencent AI Lab

Bingzhe Wu† Tencent AI Lab

Table 1: Perplexity scores of various models under diverse quantization configurations on three datasets: WikiText2 (WIKI), Pen Treebank (PT), and C4.

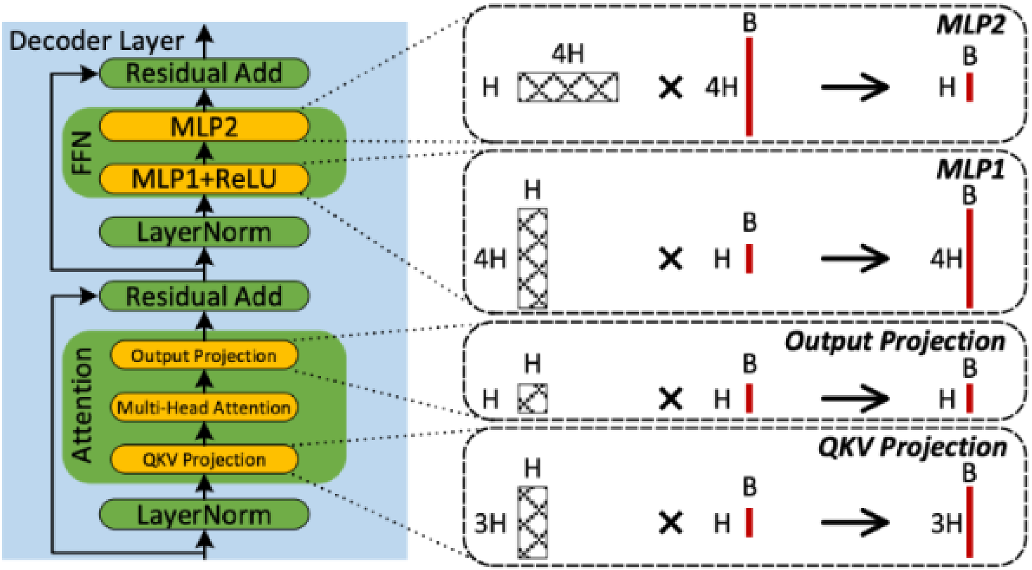
Model	OPT-1.3b			OPT-6.7b			OPT-13b			OPT-30b			OPT-66b			OPT-175b		
Task	WIKI	PT	C4	WIKI	PT	C4	WIKI	PT	C4	WIKI	PT	C4	WIKI	PT	C4	WIKI	PT	C4
FP16	14.63	16.96	14.72	10.86	13.09	11.74	10.13	12.34	11.20	9.56	11.84	10.69	9.34	11.36	10.28	8.34	12.01	10.13
W4A16	14.78	17.21	14.92	11.18	13.62	12.07	10.29	12.45	11.27	9.55	11.91	10.74	9.30	11.42	10.31	8.37	12.31	10.26
W4A8	15.39	17.79	15.48	11.21	13.74	12.11	10.90	13.40	11.62	10.22	12.41	11.01	9.46	11.73	10.57	8.43	12.24	10.49
W4A4	16.88	19.23	16.55	12.00	15.17	12.85	12.74	15.76	14.71	11.15	14.11	13.48	12.23	18.87	15.93	10.60	15.59	12.28
W4A4KV	15.26	17.65	15.37	11.26	13.44	12.03	10.59	12.80	11.54	9.99	12.18	11.01	9.75	11.64	10.61	8.40	12.38	10.54
W4A3KV	17.22	19.94	16.92	11.92	14.13	12.61	11.15	13.90	12.04	11.62	14.95	11.96	10.88	14.69	11.36	9.39	13.45	11.27
W3A3KV	18.45	21.33	18.26	12.42	14.48	13.13	11.47	14.08	12.41	11.76	14.98	12.22	11.47	15.03	11.75	10.03	13.82	11.30

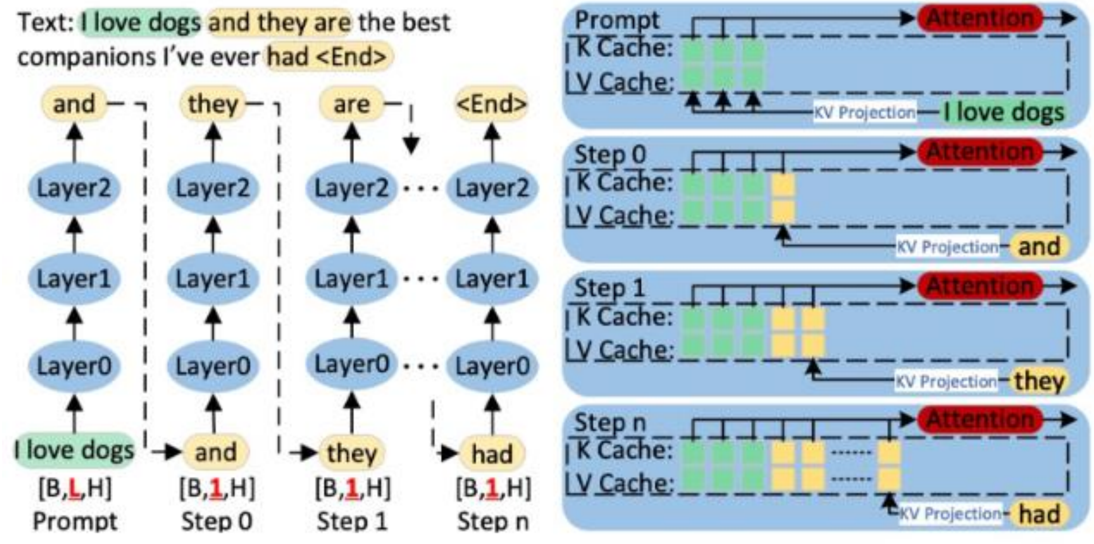
Table 3: Memory consumption (GB) of LLMs on different batch sizes and sequence lengths.

	Batch Size	1			8			64		
	Sequence Length	2048	4096	8192	2048	4096	8192	2048	4096	8192
OPT-30b	W16A16	59.4	62.3	68.1	79.7	102.9	149.3	242.0	427.5	798.6
	W4A16	17.0	19.9	25.7	37.3	60.5	106.9	199.6	385.2	756.2
	W4A8	15.6	17.1	20.1	26.0	38.0	61.8	109.5	204.9	395.7
	W4A4	14.9	15.7	17.3	20.4	26.7	39.3	64.5	114.8	215.4
	W4A4KV	15.0	15.9	17.7	21.2	28.3	42.6	71.0	127.9	241.7
	W4A3KV	14.8	15.6	17.0	19.9	25.7	37.2	60.3	106.5	198.8
	W3A3KV	11.3	12.0	13.5	16.4	22.1	33.7	56.8	102.9	195.3
OPT-66b	W16A16	128.1	133.0	142.7	162.1	200.9	278.5	433.8	744.3	1365.3
	W4A16	35.7	40.5	50.2	69.6	108.4	186.1	341.3	651.9	1272.9
	W4A8	33.3	35.8	40.7	50.6	70.5	110.1	189.5	348.1	665.4
	W4A4	32.1	33.4	36.0	41.2	51.5	72.2	113.5	196.2	361.6
	W4A4KV	32.2	33.7	36.5	42.2	53.6	76.4	122.0	213.1	395.4
	W4A3KV	32.0	33.1	35.4	39.9	49.0	67.2	103.7	176.5	322.3
	W3A3KV	24.3	25.4	27.7	32.2	41.3	59.5	96.0	168.8	314.6
OPT-175b	W16A16	335.4	344.9	363.8	401.7	477.5	629.0	932.0	1538.0	2750.1
	W4A16	91.0	100.4	119.4	157.2	233.0	384.5	687.5	1293.5	2505.6
	W4A8	86.3	91.1	100.7	119.9	158.4	235.3	389.0	696.5	1311.6
	W4A4	84.0	86.4	91.4	101.3	121.1	160.6	239.8	398.0	714.6
	W4A4KV	84.1	86.8	92.1	102.7	123.9	166.3	251.0	420.5	759.6
	W4A3KV	83.6	85.7	89.8	98.1	114.8	148.1	214.6	347.8	614.1
	W3A3KV	63.2	65.3	69.4	77.8	94.4	127.7	194.3	327.4	593.7

Contents

LLM Inference





(a) Auto-regressive Token Generation (b) KV Cache within Layer-0

- 大语言模型的部署对硬件要求较高
- 将weight和activation用低比特存储和计算能有效降低计算和内存开销
- 现有的PTQ方法一般只针对weight做4bit量化，activation仍保持8/16bit
- 激活值中存在一些异常值(outlier)
- 不同通道中的激活值取值范围差距较大

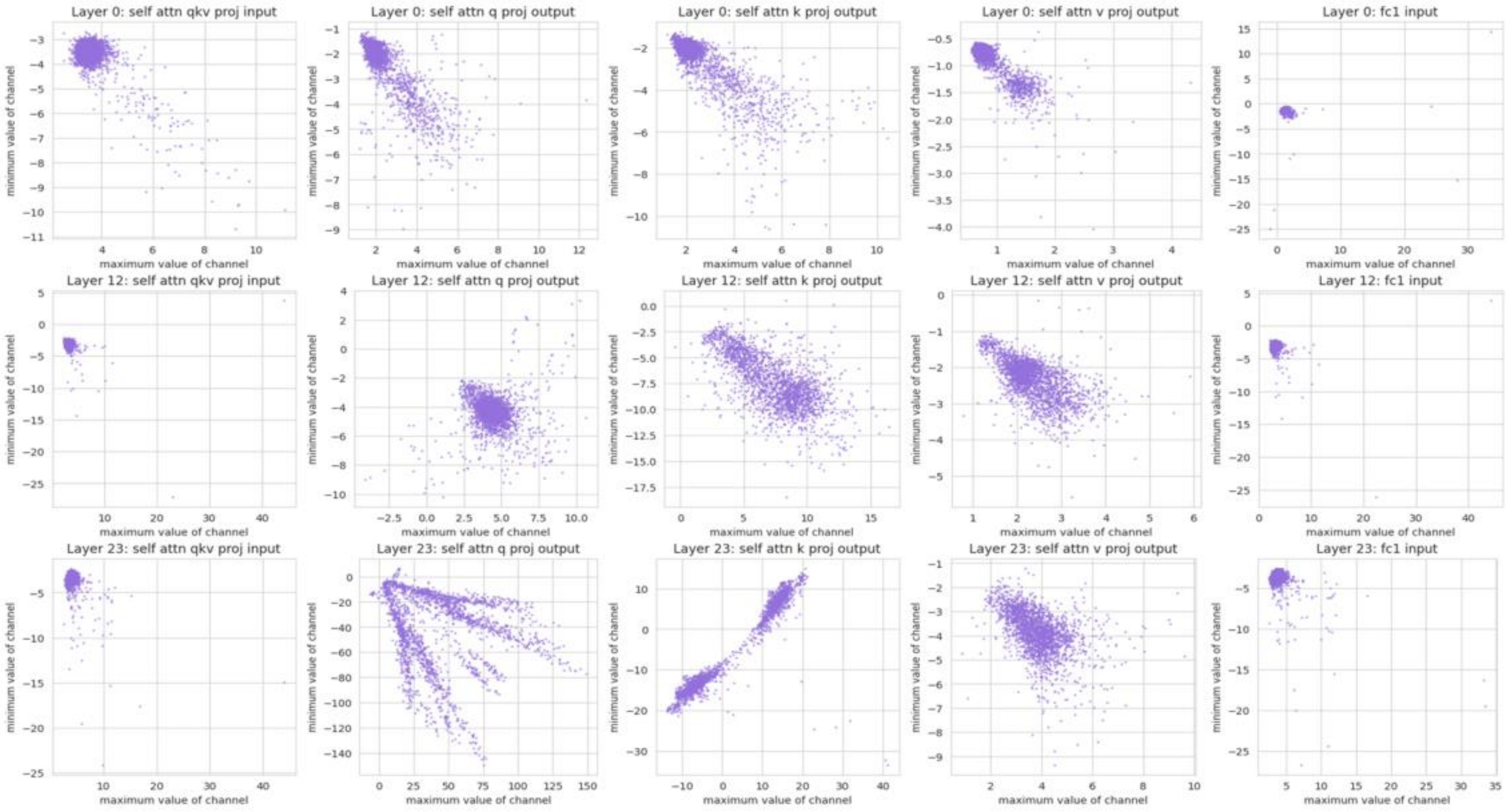


Figure 1: Demonstration of the distribution of different channels in OPT decoder layers. Each point is (maximum value, minimum value) of a channel in the activation.

Contents

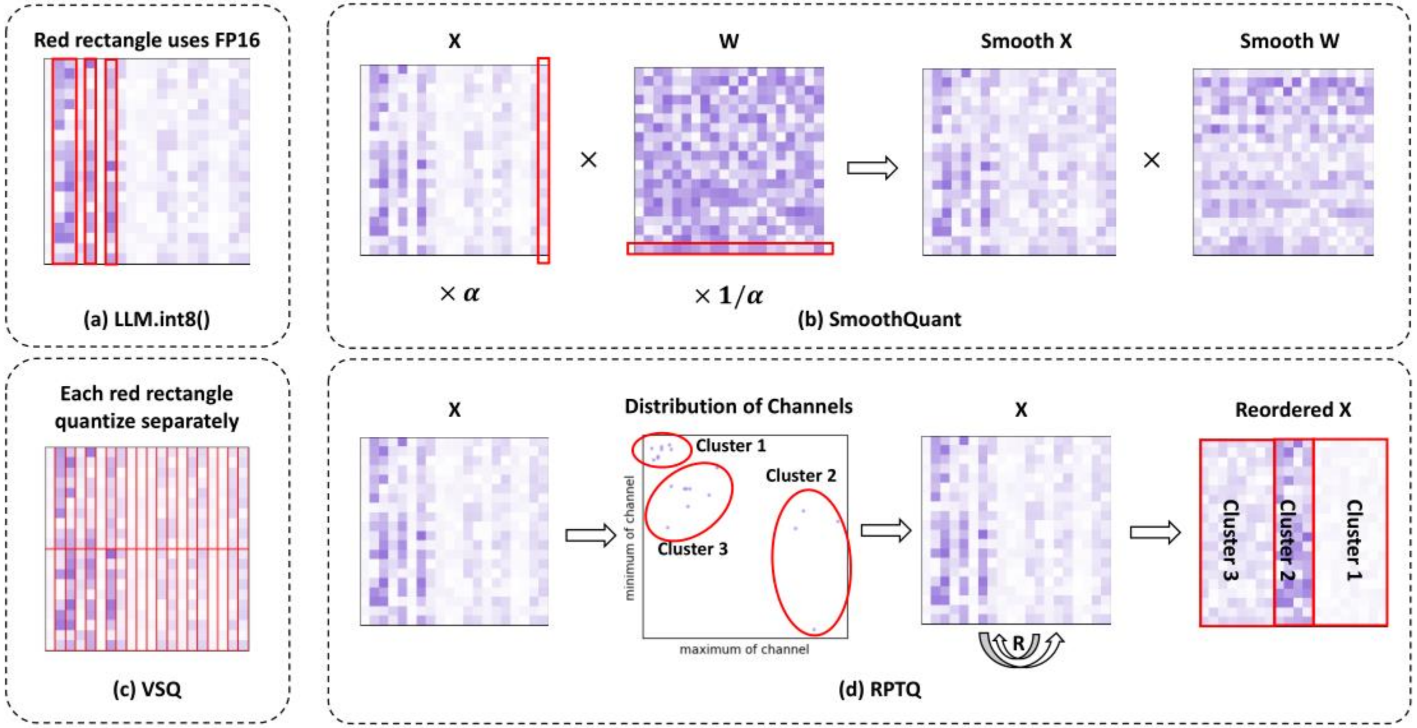


Figure 2: Demonstration of different methods to address the problem in quantizing activations.

- LLM.int8()采用FP16保存outlier，其余量化为int8
- SmoothQuant通过引入α平衡来activation和weight的量化
- VSQ对相邻的n行或n列采用相同的量化参数
- RPTQ采用先聚类再量化

- 对激活值采用非对称量化(asymmetric quantization)

$$x_q = Q_k(x, s, z) = \text{clamp}(\text{round}(\frac{x}{s}) + z, -2^{k-1}, 2^{k-1} - 1)$$

- 采用Min-Max确定scale和zero point

$$s = \frac{X_{max} - X_{min}}{2^k}, \quad z = -round(\frac{X_{max} + X_{min}}{2s}).$$

- 对activation先聚类再量化

$$X_{\min} = \min_{n=1}^{N} \min_{b=1}^{B} X_{b,n}, \quad X_{\max} = \max_{n=1}^{N} \max_{b=1}^{B} X_{b,n}.$$

根据每个channel的(min, max)使用K-means聚类

根据聚类结果对X的channel重排序，得到新的X

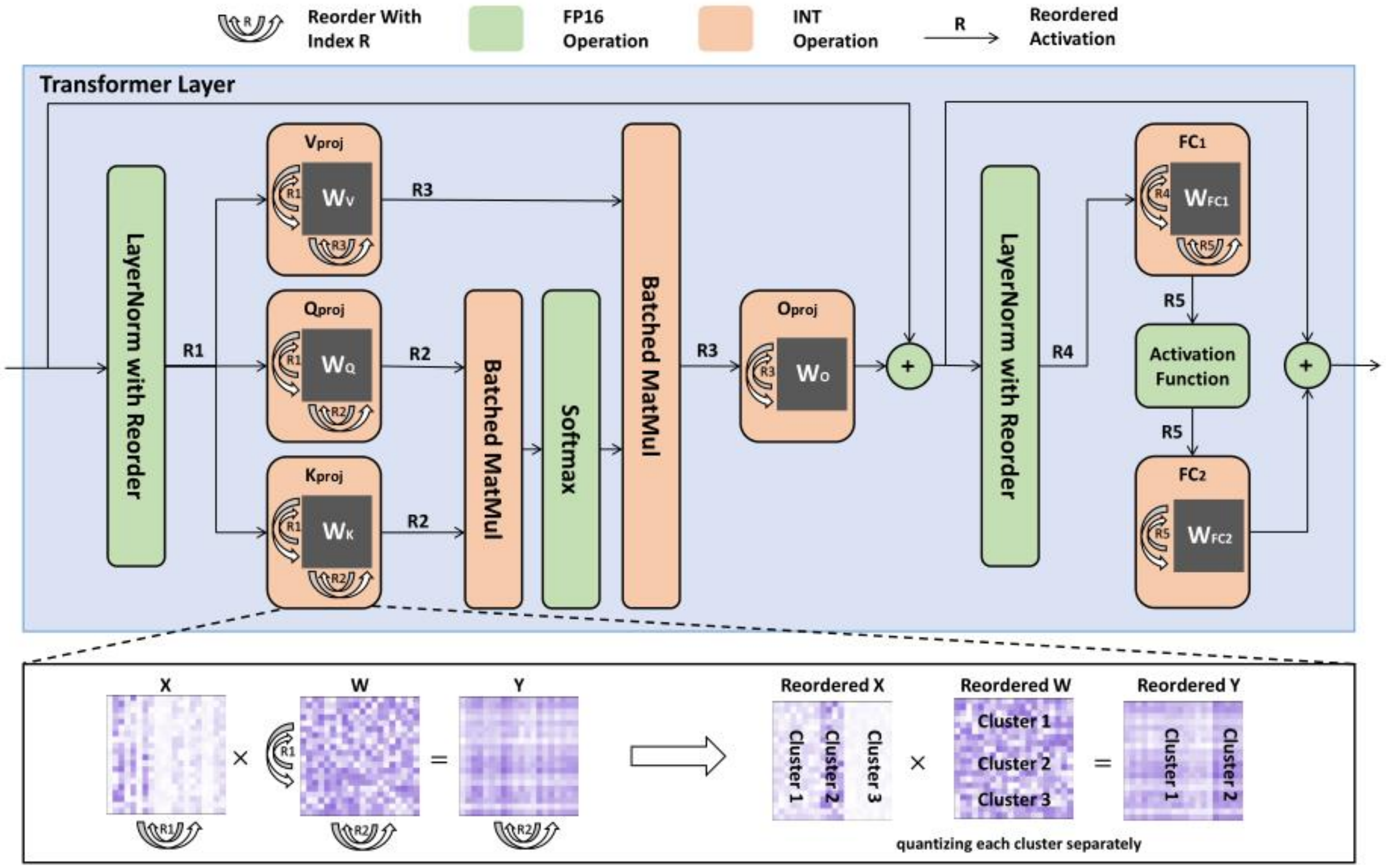


Figure 3: An overview of the inference process for a quantized transformer layer with reordered weight and activation. The reordering indexes are represented by the symbols R1 to R5. Below the figure illustrates how the weights and activations are reordered in a linear layer, where points with darker colors represent larger values.

- 重排序的结果会引起额外的开销

  1. 将重排序的过程融合到layer norm中，在写回内存时根据重排序的结果改变寻址的偏移量（running time）
  2. 根据重排序的结果直接调整权重的分布(before deployment)

- 不能破坏transformer中的残差连接

  对projection和mlp2的output channel不做重排序

- attention机制需要确保Q,K在计算聚类时要保持一致

  对Q, K的聚类按照4个值进行 $(X^{Q}_{\max,i}, X^{Q}_{\min,i}, X^{K}_{\max,i}, X^{K}_{\min,i})$

Contents

规格严格 功夫到家
1920 —— 2020

- Background & Motivation
- Method
- Evaluation
- Summary

- 对每个self-attention的头都单独聚类分析
- 从WikiText2、Pen Treebank和C4中随机抽取256条数据用作聚类的校对集
- 对OPT不同大小的模型分别测试困惑度（越小越好）
- 使用GPTQ对weight量化

Table 1: Perplexity scores of various models under diverse quantization configurations on three datasets: WikiText2 (WIKI), Pen Treebank (PT), and C4.

Model	OPT-1.3b			OPT-6.7b			OPT-13b			OPT-30b			OPT-66b			OPT-175b		
Task	WIKI	PT	C4	WIKI	PT	C4	WIKI	PT	C4	WIKI	PT	C4	WIKI	PT	C4	WIKI	PT	C4
FP16	14.63	16.96	14.72	10.86	13.09	11.74	10.13	12.34	11.20	9.56	11.84	10.69	9.34	11.36	10.28	8.34	12.01	10.13
W4A16	14.78	17.21	14.92	11.18	13.62	12.07	10.29	12.45	11.27	9.55	11.91	10.74	9.30	11.42	10.31	8.37	12.31	10.26
W4A8	15.39	17.79	15.48	11.21	13.74	12.11	10.90	13.40	11.62	10.22	12.41	11.01	9.46	11.73	10.57	8.43	12.24	10.49
W4A4	16.88	19.23	16.55	12.00	15.17	12.85	12.74	15.76	14.71	11.15	14.11	13.48	12.23	18.87	15.93	10.60	15.59	12.28
W4A4KV	15.26	17.65	15.37	11.26	13.44	12.03	10.59	12.80	11.54	9.99	12.18	11.01	9.75	11.64	10.61	8.40	12.38	10.54
W4A3KV	17.22	19.94	16.92	11.92	14.13	12.61	11.15	13.90	12.04	11.62	14.95	11.96	10.88	14.69	11.36	9.39	13.45	11.27
W3A3KV	18.45	21.33	18.26	12.42	14.48	13.13	11.47	14.08	12.41	11.76	14.98	12.22	11.47	15.03	11.75	10.03	13.82	11.30

• 在不同zero-shot任务上的实验

Table 2: Accuracy of OPT models under diverse quantization configurations on different zero-shot tasks: LAMBADA(OpenAI), PIQA, ARC(Easy), ARC(Challenge), OpenBookQA, BoolQ.

Task	LAMBADA(OpenAI) [26]					PIQA [31]				
Model	1.3b	6.7b	13b	30b	66b	1.3b	6.7b	13b	30b	66b
FP16	57.98%	61.84%	68.60%	71.41%	67.14%	72.47%	74.53%	76.87%	78.01%	78.12%
W4A16	57.46%	60.78%	68.50%	71.37%	67.06%	71.59%	74.80%	76.93%	78.29%	78.18%
W4A8	52.39%	67.35%	62.44%	64.99%	67.02%	69.69%	75.89%	75.46%	76.93%	77.52%
W4A4	49.34%	64.93%	60.23%	63.92%	68.50%	68.66%	75.40%	73.55%	76.16%	77.14%
W4A4KV	52.90%	67.39%	62.77%	64.89%	69.99%	69.26%	76.00%	74.42%	76.65%	76.98%
W4A3KV	47.02%	64.97%	61.05%	59.20%	66.23%	68.22%	75.73%	73.23%	67.46%	74.21%
W3A3KV	42.84%	64.11%	60.02%	58.33%	65.28%	68.22%	74.64%	74.10%	67.51%	75.13%

Task	ARC(Easy) [7]					ARC(Challenge) [7]				
Model	1.3b	6.7b	13b	30b	66b	1.3b	6.7b	13b	30b	66b
FP16	51.05%	58.03%	61.91%	65.31%	64.68%	29.69%	33.61%	35.66%	38.05%	38.99%
W4A16	51.17%	57.02%	61.82%	65.10%	64.89%	30.03%	32.59%	35.49%	37.96%	38.99%
W4A8	48.35%	60.18%	60.94%	63.46%	64.60%	26.36%	34.04%	35.58%	37.45%	38.82%
W4A4	47.55%	56.90%	58.41%	62.12%	63.76%	25.85%	34.30%	33.95%	36.17%	37.20%
W4A4KV	47.76%	57.74%	58.54%	63.59%	63.67%	27.64%	33.95%	34.21%	37.37%	37.71%
W4A3KV	46.29%	56.69%	56.10%	48.44%	59.00%	26.02%	33.95%	33.95%	30.71%	36.77%
W3A3KV	44.02%	55.59%	53.74%	50.42%	57.65%	26.53%	32.16%	32.50%	30.71%	34.98%

- 不同batch size和sequence length下的内存占比

Table 3: Memory consumption (GB) of LLMs on different batch sizes and sequence lengths.

	Batch Size	1			8			64		
	Sequence Length	2048	4096	8192	2048	4096	8192	2048	4096	8192
OPT-30b	W16A16	59.4	62.3	68.1	79.7	102.9	149.3	242.0	427.5	798.6
	W4A16	17.0	19.9	25.7	37.3	60.5	106.9	199.6	385.2	756.2
	W4A8	15.6	17.1	20.1	26.0	38.0	61.8	109.5	204.9	395.7
	W4A4	14.9	15.7	17.3	20.4	26.7	39.3	64.5	114.8	215.4
	W4A4KV	15.0	15.9	17.7	21.2	28.3	42.6	71.0	127.9	241.7
	W4A3KV	14.8	15.6	17.0	19.9	25.7	37.2	60.3	106.5	198.8
	W3A3KV	11.3	12.0	13.5	16.4	22.1	33.7	56.8	102.9	195.3
OPT-66b	W16A16	128.1	133.0	142.7	162.1	200.9	278.5	433.8	744.3	1365.3
	W4A16	35.7	40.5	50.2	69.6	108.4	186.1	341.3	651.9	1272.9
	W4A8	33.3	35.8	40.7	50.6	70.5	110.1	189.5	348.1	665.4
	W4A4	32.1	33.4	36.0	41.2	51.5	72.2	113.5	196.2	361.6
	W4A4KV	32.2	33.7	36.5	42.2	53.6	76.4	122.0	213.1	395.4
	W4A3KV	32.0	33.1	35.4	39.9	49.0	67.2	103.7	176.5	322.3
	W3A3KV	24.3	25.4	27.7	32.2	41.3	59.5	96.0	168.8	314.6
OPT-175b	W16A16	335.4	344.9	363.8	401.7	477.5	629.0	932.0	1538.0	2750.1
	W4A16	91.0	100.4	119.4	157.2	233.0	384.5	687.5	1293.5	2505.6
	W4A8	86.3	91.1	100.7	119.9	158.4	235.3	389.0	696.5	1311.6
	W4A4	84.0	86.4	91.4	101.3	121.1	160.6	239.8	398.0	714.6
	W4A4KV	84.1	86.8	92.1	102.7	123.9	166.3	251.0	420.5	759.6
	W4A3KV	83.6	85.7	89.8	98.1	114.8	148.1	214.6	347.8	614.1
	W3A3KV	63.2	65.3	69.4	77.8	94.4	127.7	194.3	327.4	593.7

- 不同部分在推理时的内存占比，batch size越大，K/V占比越高

Table 6: The memory proportion of different parts in LLMs.

Batch Size		1						64					
Sequence Length		2048			8192			2048			8192		
Model	Precision	Weight	K/V	Dynamic	Weight	K/V	Dynamic	Weight	K/V	Dynamic	Weight	K/V	Dynamic
OPT-1.3b	FP16	85.35%	12.12%	2.53%	59.30%	33.67%	7.03%	8.35%	75.83%	15.82%	2.23%	80.89%	16.88%
	W4A16	59.30%	33.67%	7.03%	26.70%	60.65%	12.65%	2.23%	80.89%	16.88%	0.57%	82.27%	17.17%
	W4A8	73.48%	20.86%	5.66%	40.92%	46.47%	12.62%	4.15%	75.39%	20.47%	1.07%	77.81%	21.12%
	W4A4	83.45%	11.85%	4.70%	55.76%	31.66%	12.57%	7.30%	66.35%	26.35%	1.93%	70.19%	27.88%
	W4A4KV	80.47%	11.42%	8.11%	50.74%	28.81%	20.45%	6.05%	54.95%	39.00%	1.58%	57.57%	40.85%
	W4A3KV	82.94%	8.83%	8.23%	54.86%	23.36%	21.78%	7.06%	48.10%	44.84%	1.86%	50.79%	47.35%
	W3A3KV	78.47%	11.14%	10.39%	47.68%	27.08%	25.24%	5.39%	48.96%	45.65%	1.40%	51.03%	47.57%
OPT-6.7b	FP16	91.70%	7.17%	1.12%	73.43%	22.98%	3.59%	14.73%	73.74%	11.53%	4.14%	82.90%	12.96%
	W4A16	73.43%	22.98%	3.59%	40.86%	51.14%	8.00%	4.14%	82.90%	12.96%	1.07%	85.55%	13.38%
	W4A8	84.16%	13.17%	2.68%	57.04%	35.70%	7.26%	7.66%	76.73%	15.60%	2.03%	81.41%	16.56%
	W4A4	90.79%	7.10%	2.11%	71.13%	22.26%	6.62%	13.34%	66.80%	19.86%	3.71%	74.22%	22.07%
	W4A4KV	89.30%	6.99%	3.71%	67.60%	21.15%	11.25%	11.54%	57.75%	30.71%	3.16%	63.22%	33.62%
	W4A3KV	90.94%	5.34%	3.73%	71.50%	16.78%	11.72%	13.55%	50.89%	35.55%	3.77%	56.65%	39.58%
	W3A3KV	88.27%	6.91%	4.82%	65.30%	20.43%	14.27%	10.52%	52.68%	36.80%	2.86%	57.19%	39.95%
OPT-13b	FP16	93.28%	5.97%	0.75%	77.64%	19.87%	2.49%	17.83%	73.03%	9.14%	5.15%	84.31%	10.55%
	W4A16	77.64%	19.87%	2.49%	46.47%	47.58%	5.95%	5.15%	84.31%	10.55%	1.34%	87.69%	10.97%
	W4A8	87.05%	11.14%	1.81%	62.69%	32.09%	5.22%	9.50%	77.83%	12.66%	2.56%	83.81%	13.64%
	W4A4	92.66%	5.93%	1.41%	75.94%	19.44%	4.62%	16.47%	67.47%	16.05%	4.70%	76.99%	18.31%
	W4A4KV	91.64%	5.86%	2.49%	73.27%	18.75%	7.98%	14.62%	59.90%	25.48%	4.11%	67.28%	28.62%
	W4A3KV	93.04%	4.47%	2.50%	76.97%	14.78%	8.26%	17.28%	53.07%	29.66%	4.96%	60.97%	34.07%
	W3A3KV	90.93%	5.82%	3.25%	71.48%	18.30%	10.23%	13.54%	55.46%	31.00%	3.77%	61.73%	34.50%
OPT-30b	FP16	95.12%	4.42%	0.46%	82.97%	15.42%	1.61%	23.34%	69.42%	7.24%	7.07%	84.15%	8.77%
	W4A16	82.97%	15.42%	1.61%	54.92%	40.83%	4.26%	7.07%	84.15%	8.77%	1.87%	88.87%	9.26%
	W4A8	90.45%	8.41%	1.14%	70.32%	26.14%	3.54%	12.90%	76.70%	10.40%	3.57%	84.92%	11.51%
	W4A4	94.73%	4.40%	0.87%	81.79%	15.20%	3.01%	21.91%	65.17%	12.92%	6.56%	77.98%	15.46%
	W4A4KV	94.08%	4.37%	1.55%	79.89%	14.85%	5.26%	19.89%	59.14%	20.97%	5.84%	69.51%	24.64%
	W4A3KV	95.14%	3.32%	1.54%	83.04%	11.57%	5.39%	23.43%	52.24%	24.33%	7.10%	63.38%	29.52%
	W3A3KV	93.62%	4.35%	2.03%	78.59%	14.61%	6.80%	18.66%	55.49%	25.84%	5.42%	64.53%	30.05%

- 聚类个数对困惑度的影响，理论上越多越好

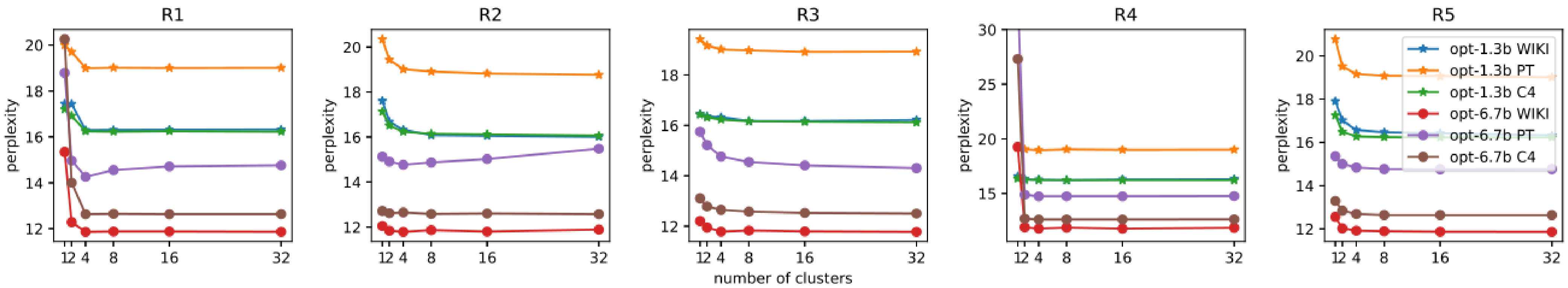


Figure 4: The ablation study to evaluate the performance of the clustering method under the W16A4 configuration. We tested different numbers of clusters (1, 2, 4, 8, and 32) for R1 to R5.

Contents

- 第一篇将大语言模型的activation量化到3/4bit的论文
- 与GPTQ结合，实现了W4A4的量化（一般都是W4A16）
- Insight：对activation按照channel先聚类再在簇内共享量化的参数
- 相较于常规的对称量化，采用了非对称量化的模式
- 在batch size很大时，能显著降低KV cache的内存占用