LLM的量化压缩进展及下一步推进方向

袁之航

20230420

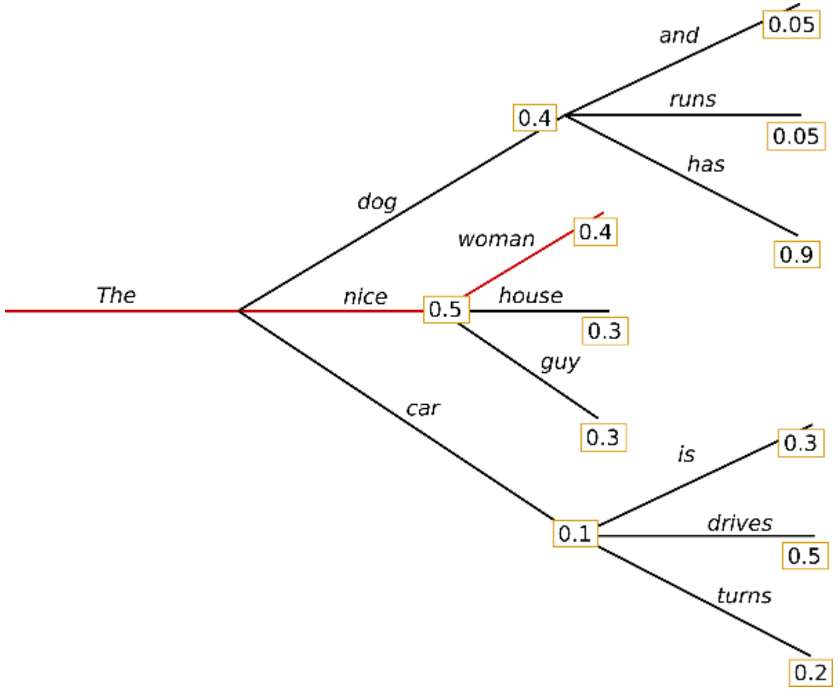
袁之航个人简介

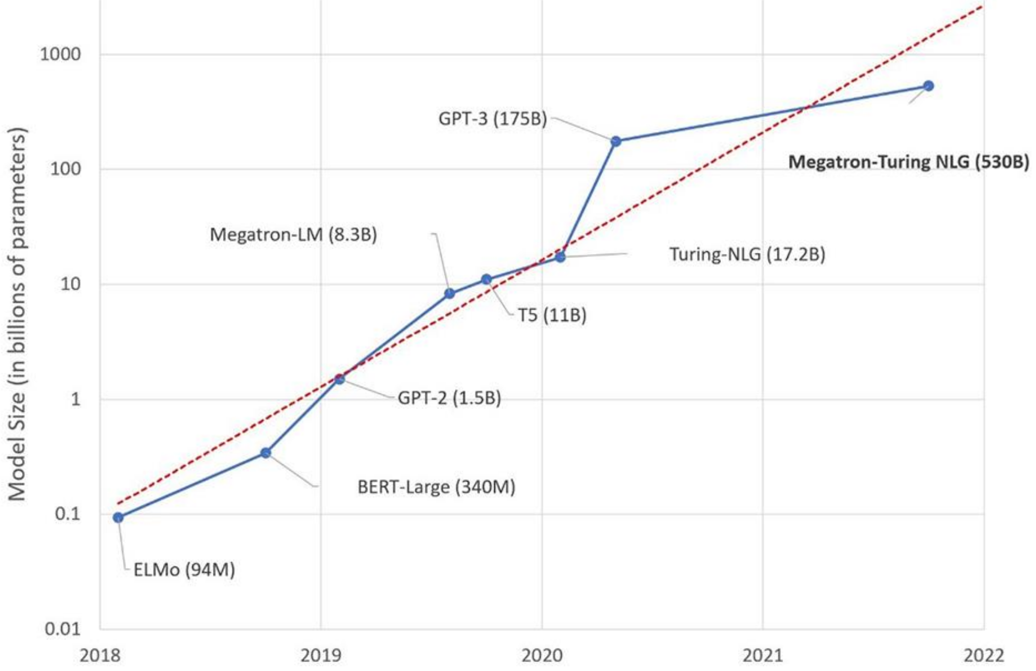
- 嗨，大家好！我于2017年获得北京大学学士学位，于2022年获得北京大学计算机学院博协士学位（系统结构方向）。
- 研究方向为神经网络的量化及推理加速、深度学习的软硬件同优化
- 发表论文十余篇，其中RPTQ是第一篇将LLM的激活量化压缩推进到3比特的工作。PTQ4DM是第一篇量化Stable Diffusion的工作。PTQ4ViT是第一篇ViT 8比特量化不掉点的工作。
- 2021年加入存算一体芯片创业公司后摩智能，参与了多款AI加速器设计，负责芯片量化方案的设计，领导量化算法和量化工具链的开发，并推进了多项研究成果落地。

Large Language Model (LLM)

- LM是语言模型，对于人类语言表达的建模
- LLM是语言大模型（参数量大、计算量大）

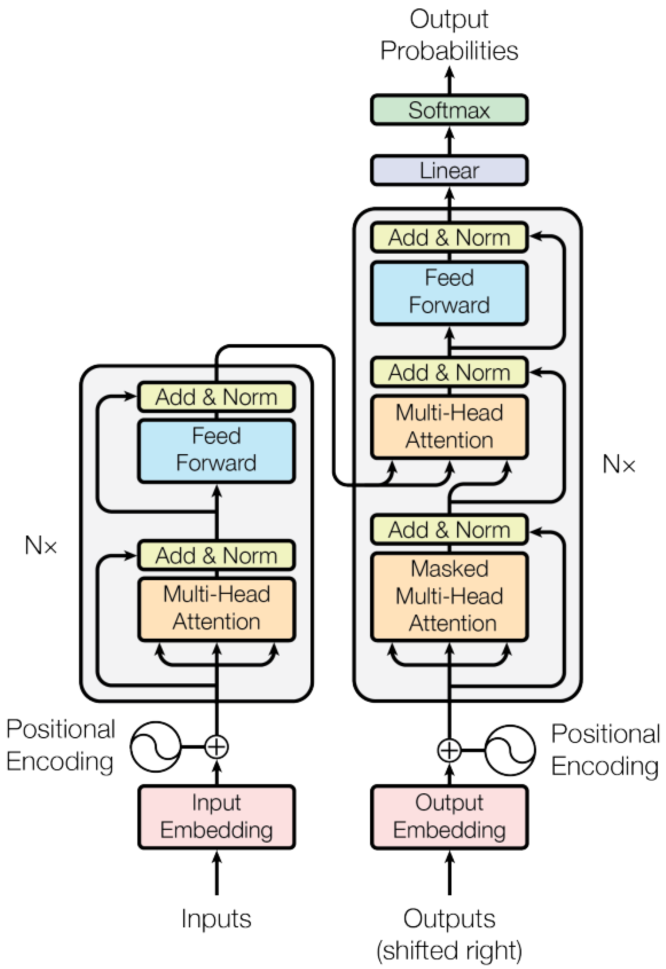
$$w_1, \dots, w_{i-1}$$



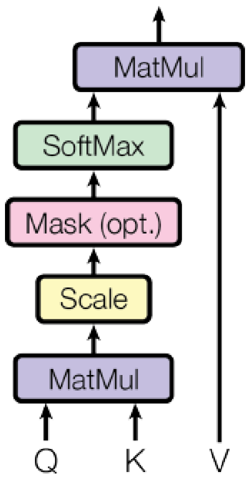


von Platen P. How to generate text: using different decoding methods for language generation with transformers[J]. Hugging Face, 2020.
Smith S, Patwary M, Norick B, et al. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model[J]. arXiv preprint arXiv:2201.11990, 2022.

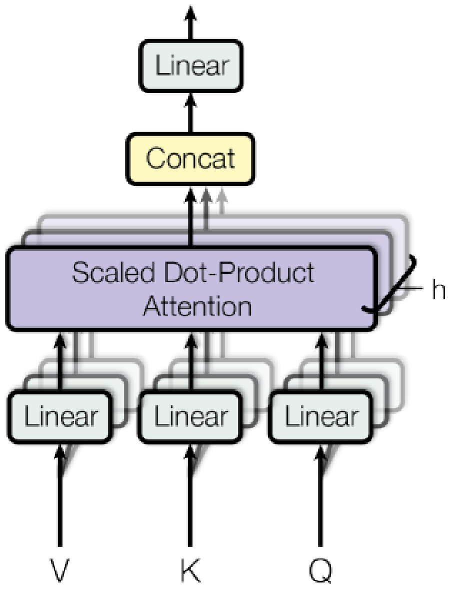
当前LLM是Transformer



Scaled Dot-Product Attention

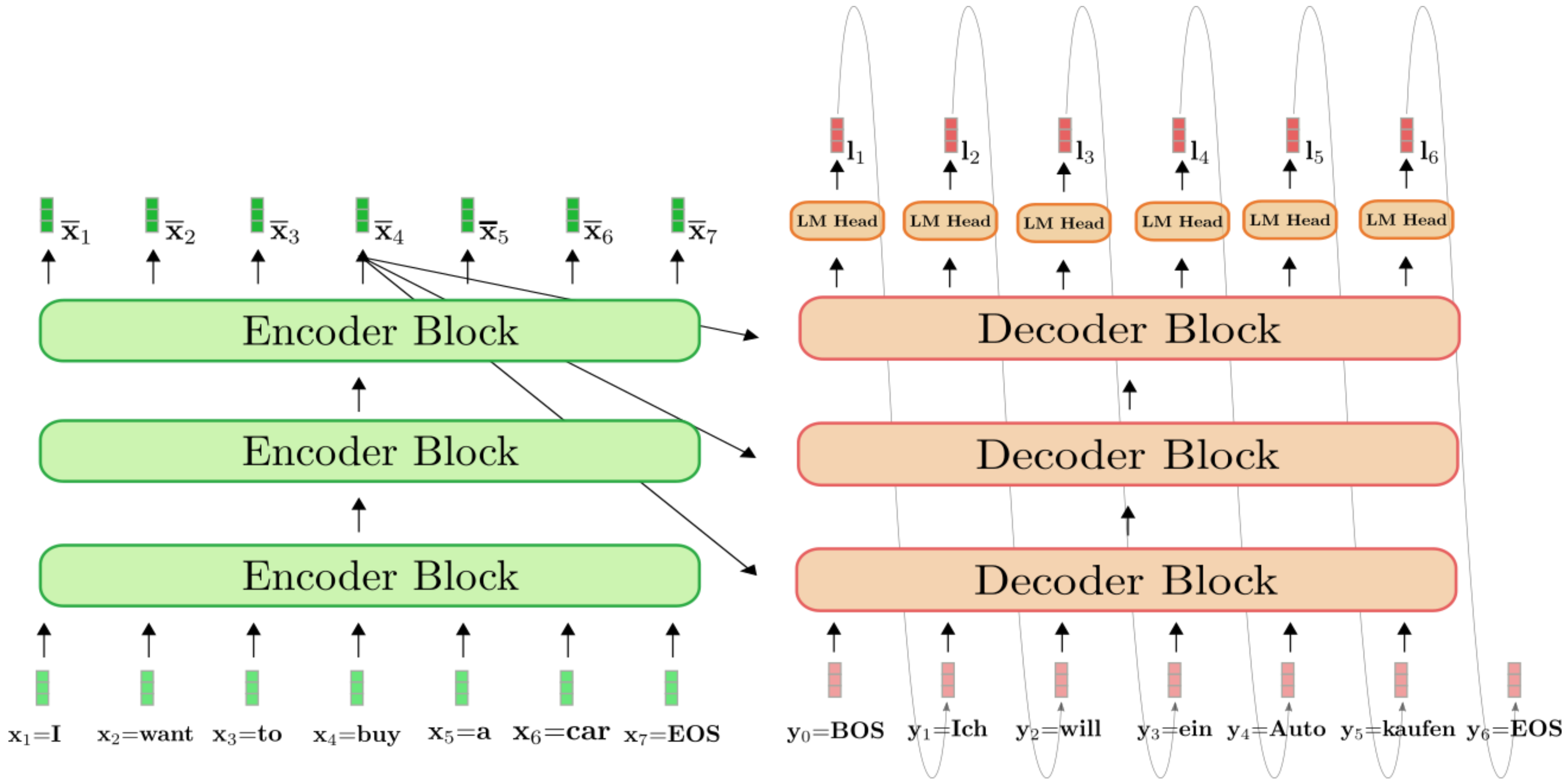


Multi-Head Attention



Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

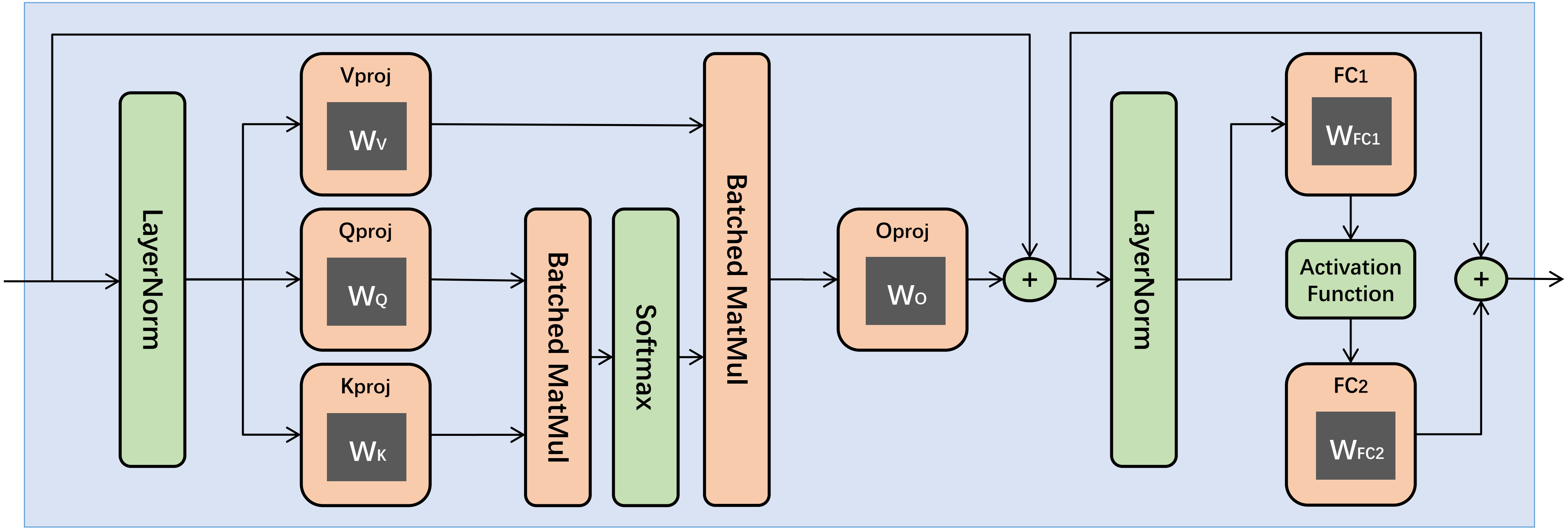
Token 生成流程



Hugging Face, The ultimate guide to Transformer-based Encoder-Decoder Models

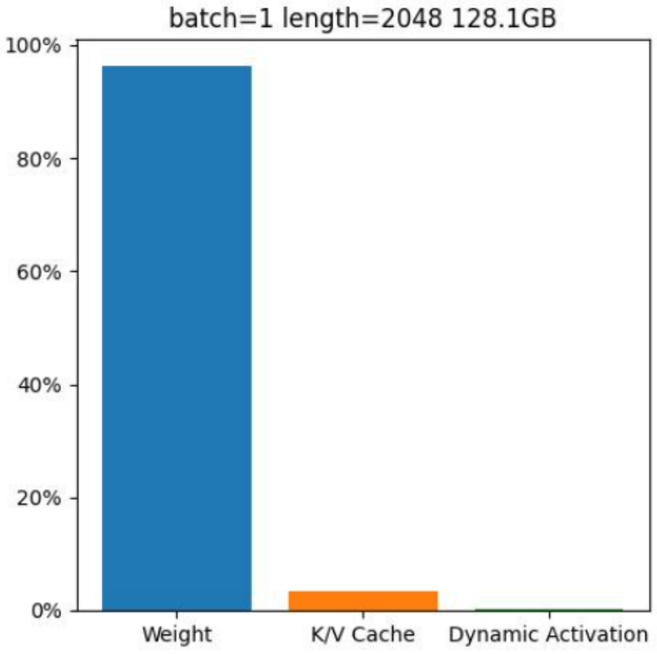
LLM的显存开销在哪里？

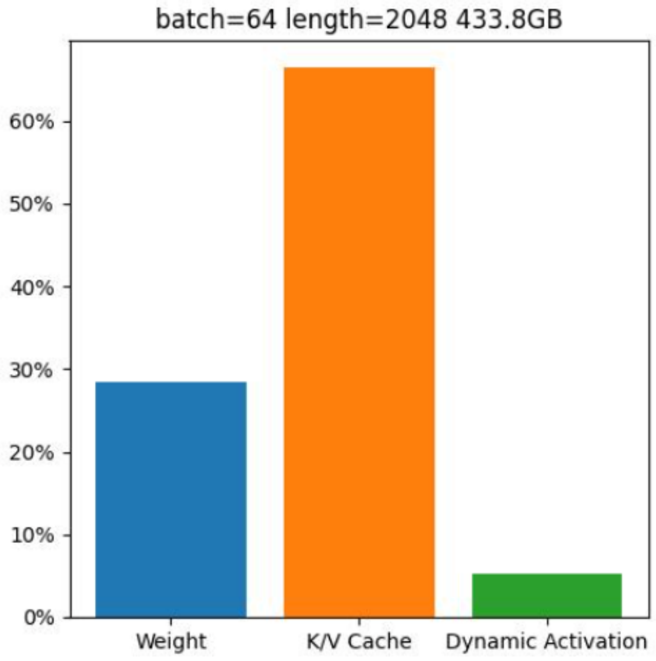
- Weight
- Activation(Key/Value Cache)
- Activation(Dynamic)

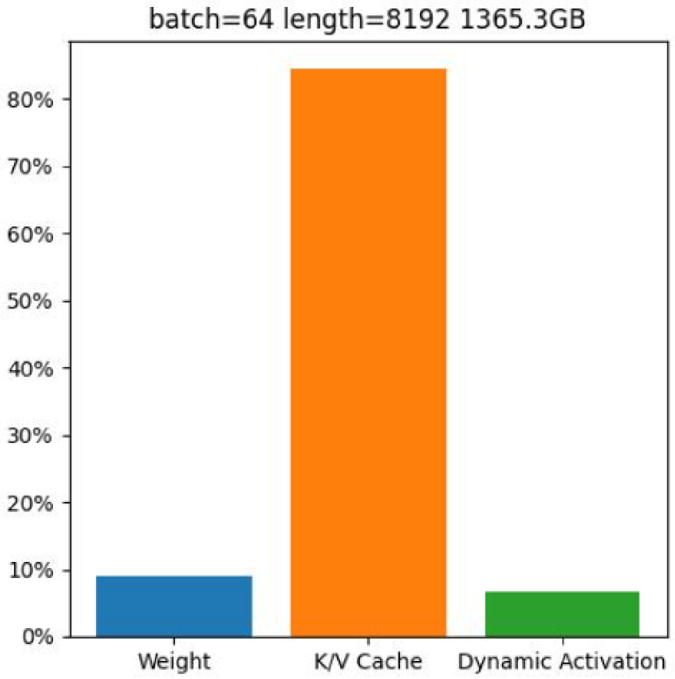


LLM的显存开销在哪里？

Dynamic Activation 可以通过融合算子大幅度降低
但K/V Cache和Weight只能通过量化等方法来缩减







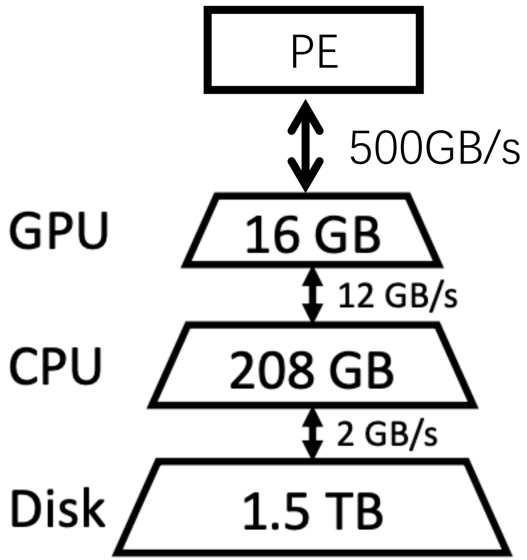
小batch小length主要是Weight

大batch、长序列会大幅增加K/V Cache的占比

LLM部署的主要问题：Memory

- 低batch：计算/访存比低，计算单元等数据
  - 每次inference有几十GB到几百GB的访存
- 高batch：显存放不下
  - 如果存储到主存上，将会完全卡在带宽

- 例子：100GB 访存
  - 如果放在GPU显存（0.2s）
  - 如果放在CPU主存（8.3s）
  - 如果放在Disk（50s）



量化压缩缓解Memory问题

	Batch Size	1			8			64		
	Sequence Length	2048	4096	8192	2048	4096	8192	2048	4096	8192
OPT-13b	W16A16	26.2	27.9	31.4	38.5	52.5	80.7	136.9	249.4	474.5
	W4A16	7.9	9.6	13.1	20.2	34.2	62.4	118.6	231.1	456.1
	W4A8	7.0	7.9	9.7	13.4	20.6	35.2	64.2	122.4	238.6
	W4A4	6.6	7.1	8.0	10.0	13.8	21.6	37.1	68.0	129.9
	W4A4KV	6.7	7.2	8.3	10.6	15.0	23.9	41.7	77.4	148.6
	W4A3KV	6.6	7.0	7.9	9.8	13.4	20.7	35.3	64.6	123.0
	W3A3KV	5.0	5.5	6.4	8.2	11.9	19.2	33.8	63.0	121.5
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

Post-training Quantization (PTQ)

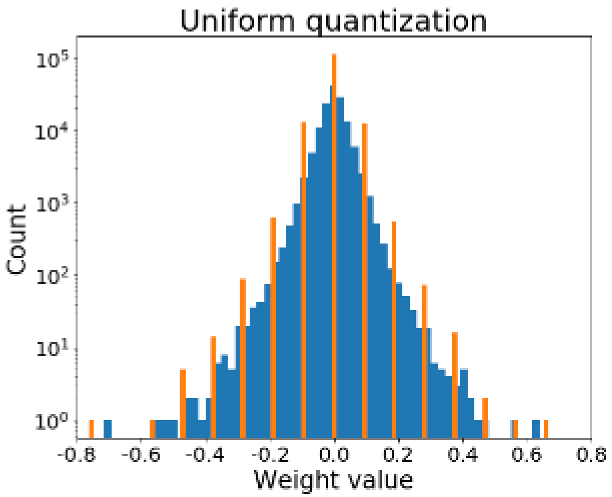
- QAT
  - 效果好
  - 需要增加训练时间(多一些training iterations)
  - 对于已经具有高训练成本的大规模语言模型（LLMs）可能负担不起

- PTQ方法
  - 较QAT精度更差
  - 不需要额外的训练，LLM几小时即可量化完成
  - 对于LLMs更可行

Uniform Quantization

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

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



对Weight的量化

- GPTQ成功推进到4bit\3bit

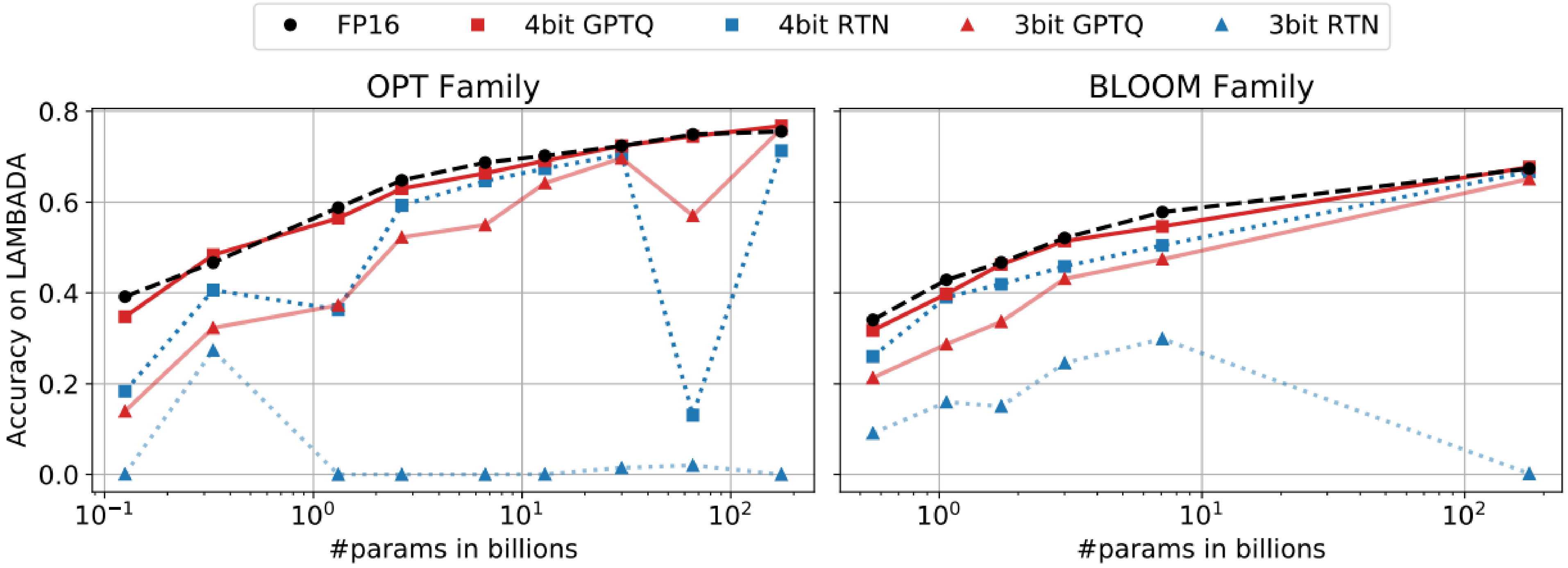
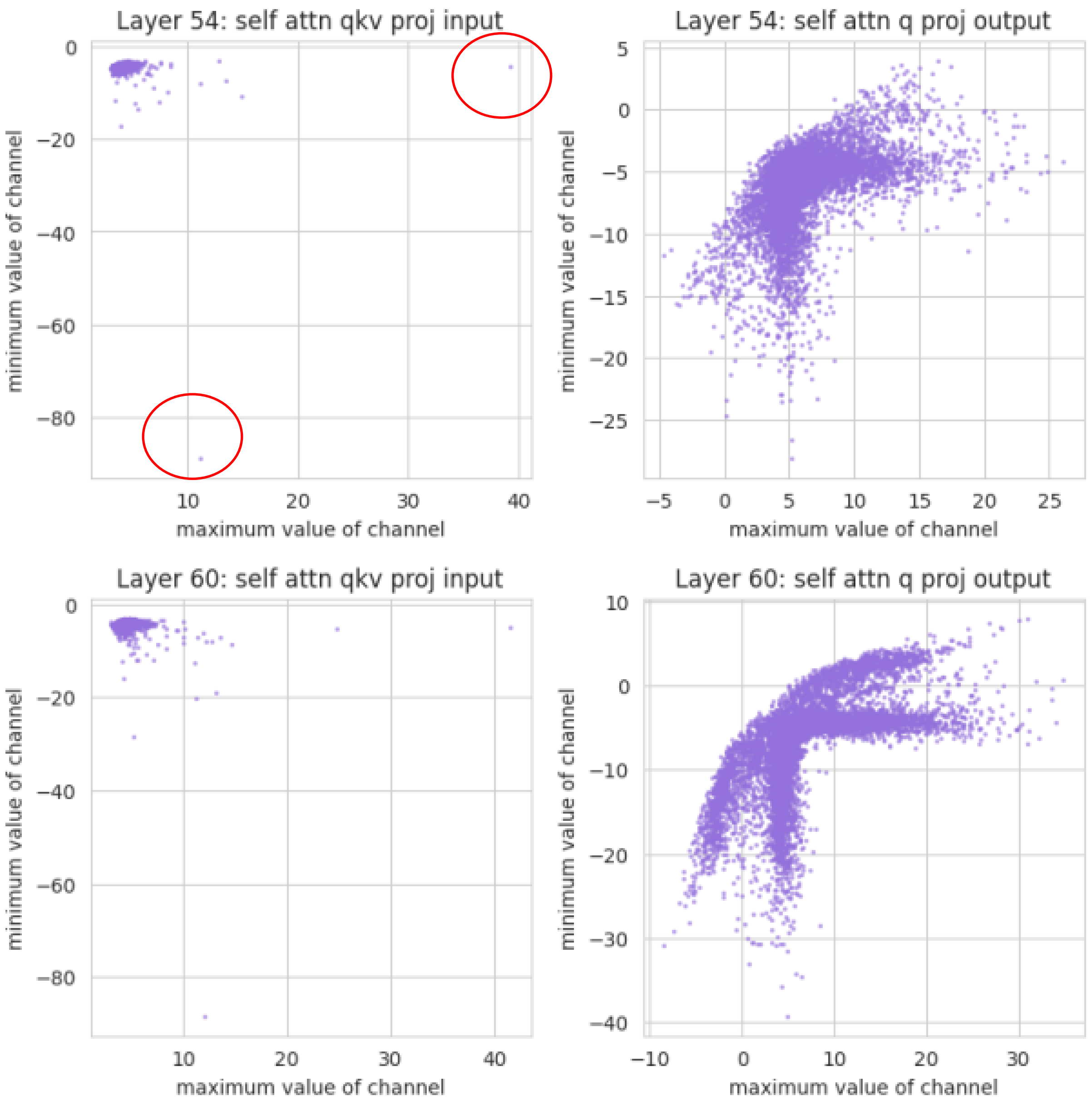


Figure 3: The accuracy of OPT and BLOOM models post-GPTQ, measured on LAMBADA.

Frantar E, Ashkboos S, Hoefler T, et al. GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers[J]. arXiv preprint arXiv:2210.17323, 2022.

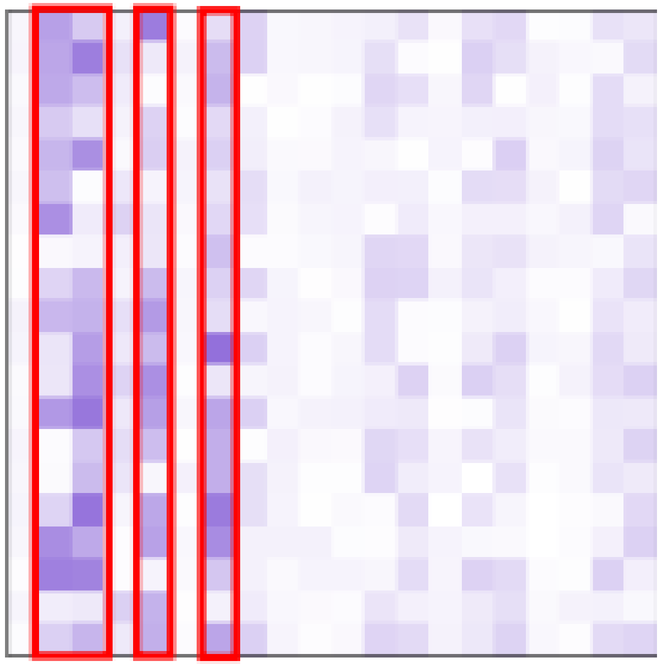
Activation量化的挑战

- 有outlier的channel
- 不同channel的range差异性



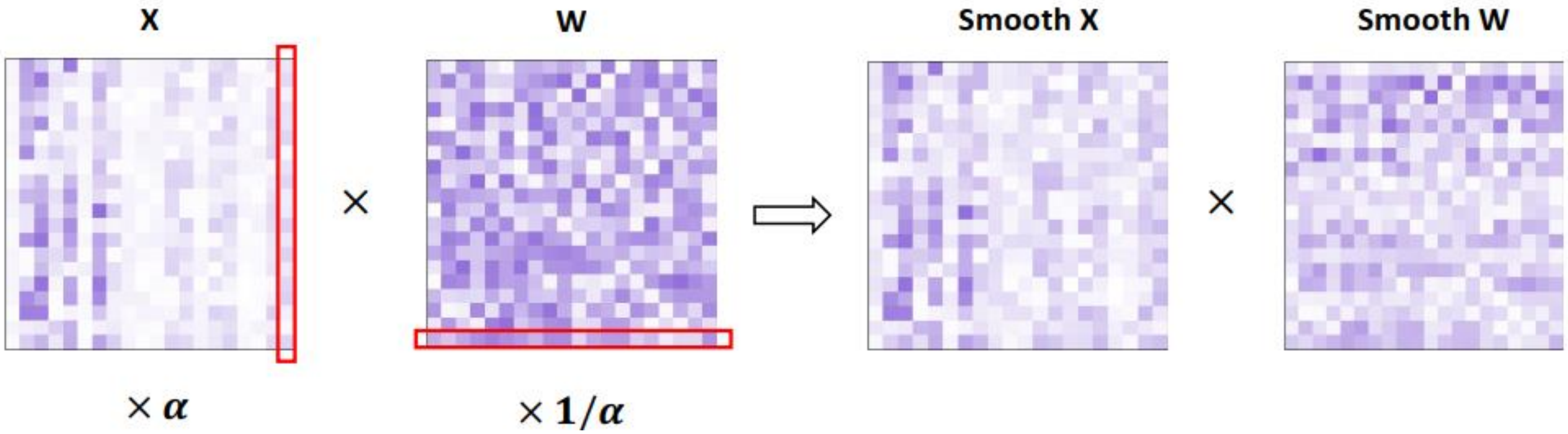
LLM.int8()

Red rectangle uses FP16



$$\mathbf{C}_{f16} \approx \sum_{h \in O} \mathbf{X}^h_{f16} \mathbf{W}^h_{f16} + \mathbf{S}_{f16} \cdot \sum_{h \notin O} \mathbf{X}^h_{i8} \mathbf{W}^h_{i8}$$

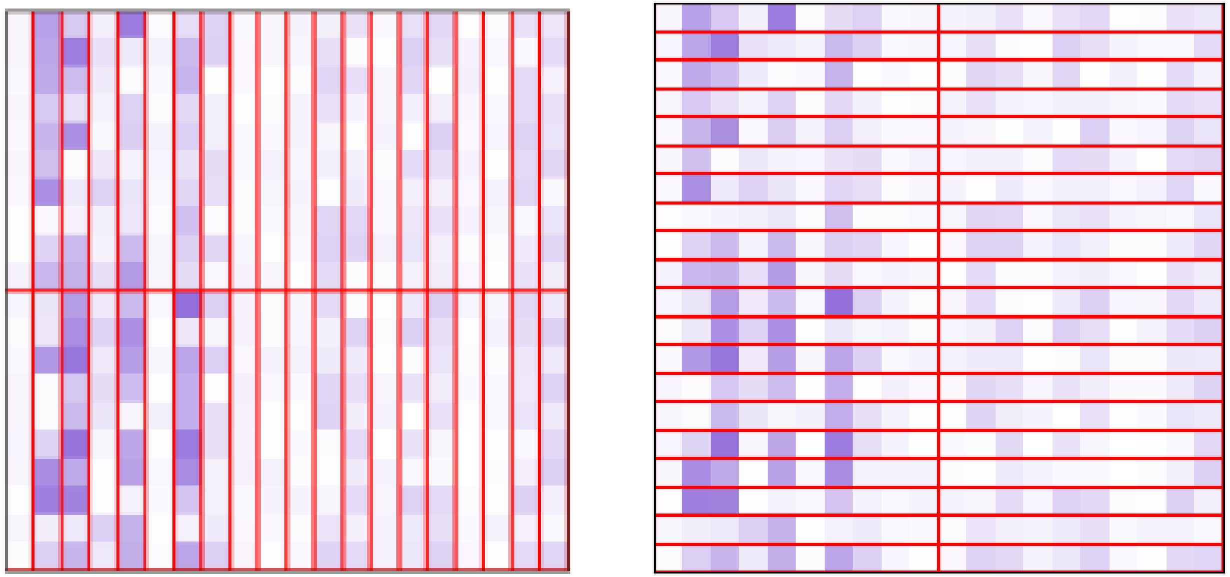
SmoothQuant



VSQ

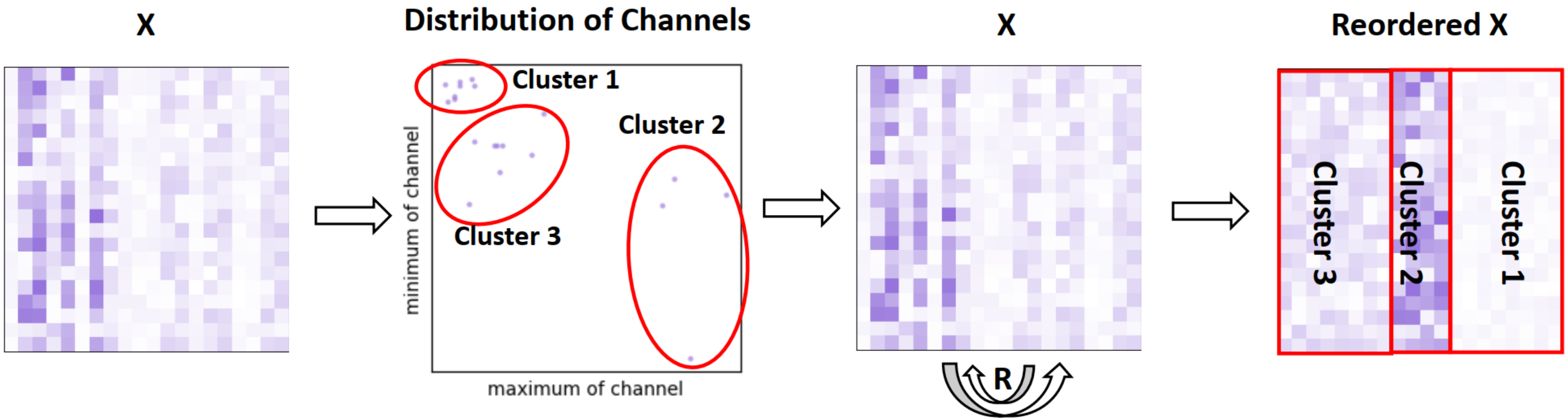
- column方向VSQ
- row方向VSQ

Each red rectangle
quantize separately

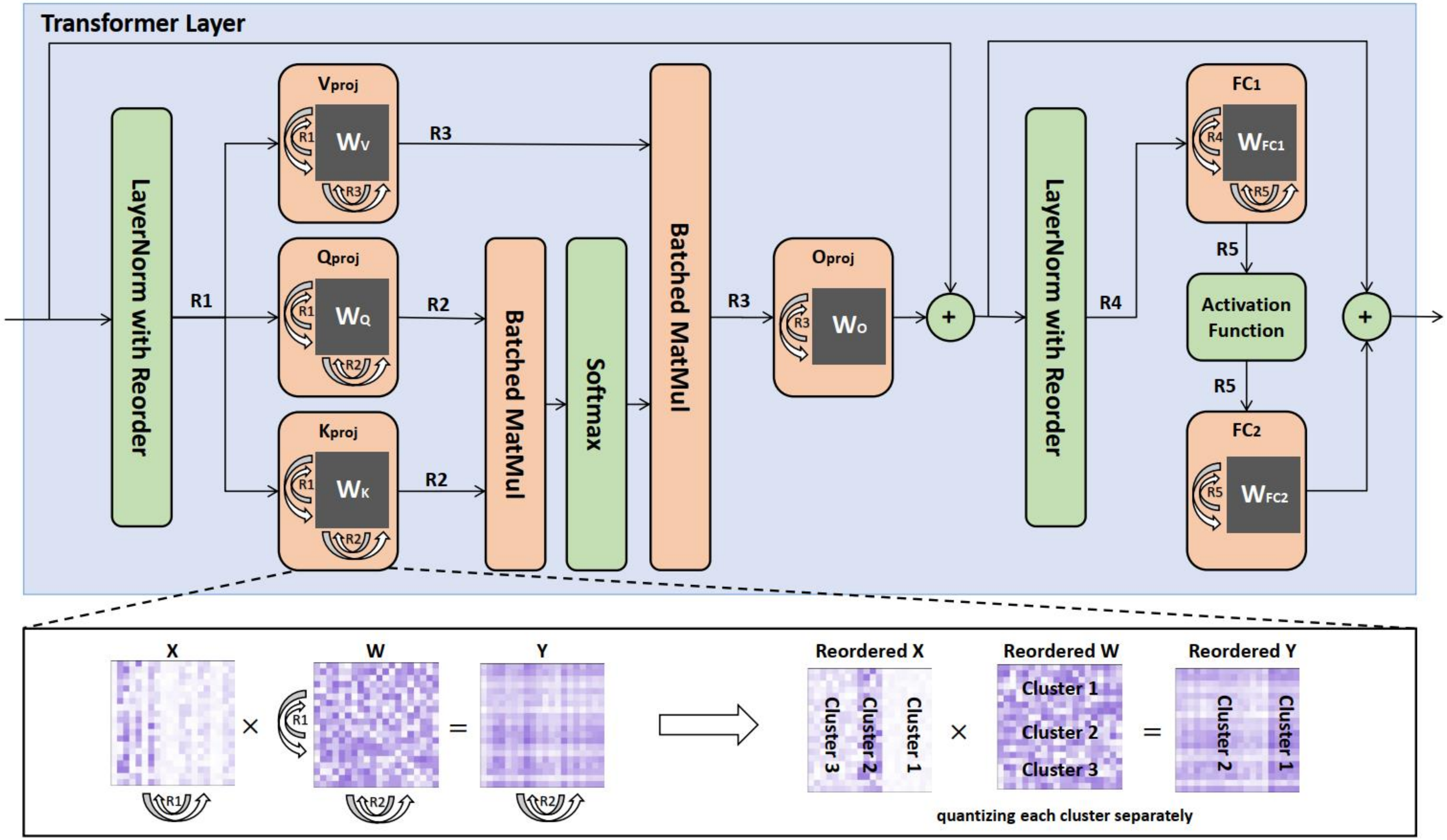


RPTQ: Clustering and Reordering

- 不同cluster的activation用不同的quantization parameters
- K-Means做聚类



Avoid Explicit Reordering and Misalignment

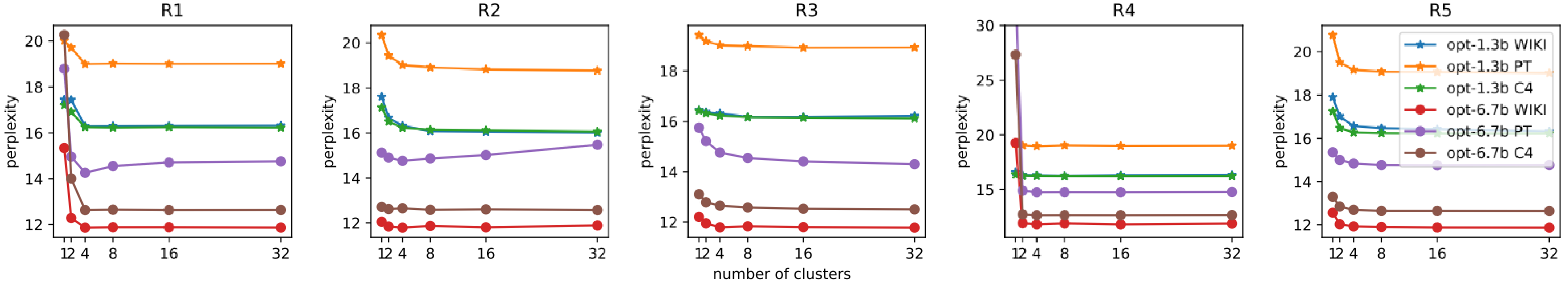


Results

Task	LAMBADA(OpenAI) [24]					PIQA [29]				
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%
Task	OpenBookQA [22]					BoolQ [6]				
Model	1.3b	6.7b	13b	30b	66b	1.3b	6.7b	13b	30b	66b
FP16	33.00%	38.00%	39.00%	40.20%	41.60%	57.73%	67.03%	65.90%	70.45%	70.85%
W4A16	31.80%	37.40%	39.20%	40.60%	42.00%	58.99%	59.72%	66.66%	70.70%	70.55%
W4A8	32.40%	38.00%	38.60%	39.40%	41.80%	46.88%	65.93%	66.57%	70.64%	71.07%
W4A4	32.60%	38.40%	38.00%	38.60%	42.00%	41.37%	65.44%	58.47%	67.70%	70.24%
W4A4KV	32.60%	38.40%	38.00%	39.80%	41.60%	43.33%	62.11%	62.47%	68.22%	70.79%
W4A3KV	32.80%	36.80%	37.00%	34.00%	39.40%	42.84%	61.31%	57.76%	61.74%	67.06%
W3A3KV	28.40%	35.20%	37.20%	32.40%	38.60%	46.23%	60.79%	65.07%	63.08%	67.49%

Ablation

- 4及以上聚类数量即可达到较好效果



下一步推进方向

- RPTQv2
  - Cluster-wise Mixed Precision
  - Dynamic Quantization for high variation channels

- QLoRA
  - Make QAT affordable for LLM

- QAT
  - Distributed Training
  - Extreme Low bit-width (3bit, ternary or binary)

谢谢

- 相信LLM的量化压缩一定是未来的主流部署技术
- 量化压缩还有大把的课题值得研究，号召大家来试试
  - 动态量化
  - sub-bit量化
  - 软硬协同优化