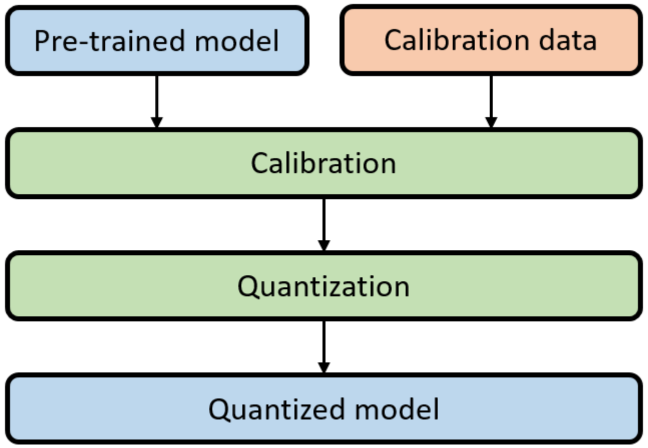
Outlier Suppression+: Accurate quantization of large language models by equivalent and optimal shifting and scaling

EMNLP 2023

Shared by: **Chao Zeng**

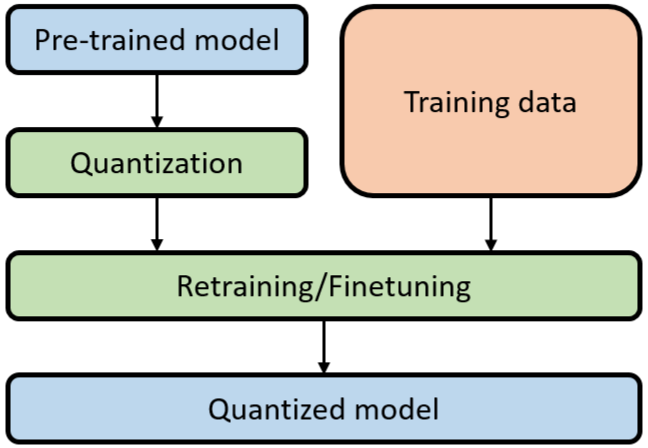
2024.01.30

模型量化：模型量化就是通过某种方法将浮点模型转为定点模型。即模型float32的权重和激活都是通过模型量化，将模型变成int8、int4等定点模型。



量化方法：
QAT (Quantization-Aware Training): 利用训练数据重新训练网络，不实用；

PTQ (Post-Training Quantization):利用校准数据获得量化校准参数（或不需要），更常用。



Weight-only quantization

AWQ (Activation-aware Weight Quantization)

AWQ认为权重并非同等重要，仅保护1%的显著权重可以大大减少量化误差。然后，我们建议通过观察激活来搜索保护显著权重的最佳通道缩放。

$$\mathbf{s} = f(\mathbf{s_X}, \mathbf{s_W}) = \mathbf{s_X}^{\alpha} \cdot \mathbf{s_W}^{-\beta}, \quad \alpha^*, \beta^* = \arg\min_{\alpha,\beta} \mathcal{L}(\mathbf{s_X}^{\alpha} \cdot \mathbf{s_W}^{-\beta}) \tag{2}$$

$$\mathbf{s}^* = \arg\min_{\mathbf{s}} \mathcal{L}(\mathbf{s}), \quad \mathcal{L}(\mathbf{s}) = \|Q(\mathbf{W} \cdot \mathbf{s})(\mathbf{s}^{-1} \cdot \mathbf{X}) - \mathbf{W}\mathbf{X}\| \tag{1}$$

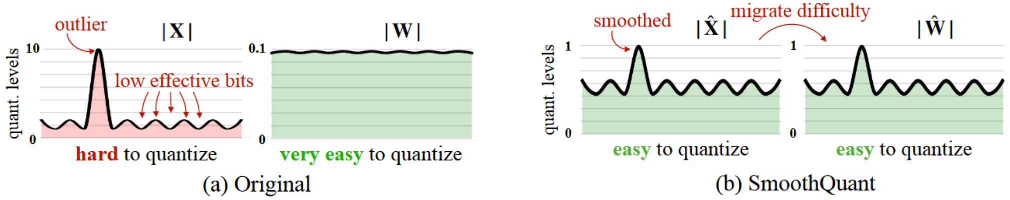
GPTQ (Generative Pretrained Transformer Quantization)

GPTQ通过估计二阶导信息估计进行量化后误差优化。

$$\mathbf{H}_{-q}^{-1} = \left(\mathbf{H}^{-1} - \frac{1}{[\mathbf{H}^{-1}]_{qq}} \mathbf{H}_{:,q}^{-1} \mathbf{H}_{q,:}^{-1}\right)_{-p}. \tag{3}$$

Weight-activation quantization

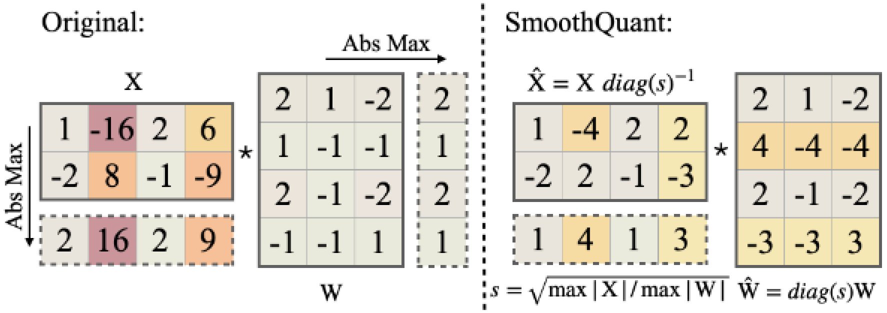
SmoothQuant



Smooth quant的思想是找到一个合适的向量scale(s)，对激活值X以及权重W做合理的缩放，将激活值一部分量化难度转移到权重上，让激活值更加容易量化。

$$\mathbf{Y} = (\mathbf{X}\mathrm{diag}(\mathbf{s})^{-1}) \cdot (\mathrm{diag}(\mathbf{s})\mathbf{W}) = \hat{\mathbf{X}}\hat{\mathbf{W}} \quad (3)$$

$$\mathbf{s}_j = \max(|\mathbf{X}_j|)^{\alpha} / \max(|\mathbf{W}_j|)^{1-\alpha} \quad (4)$$



FPTQ (Fine-grained Post-Training Quantization)

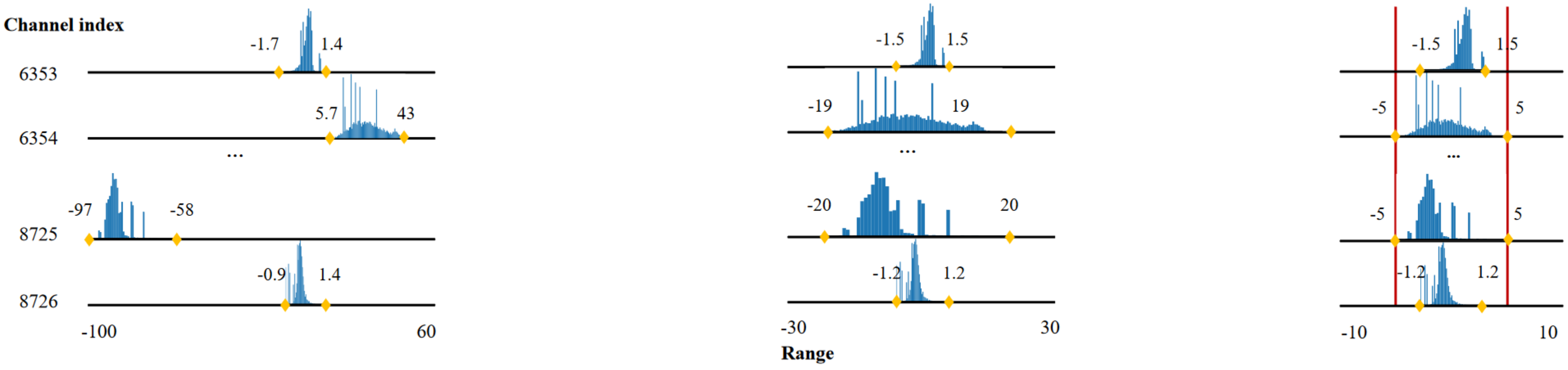
FPTQ针对activation和weight之间的关系选择更加细粒度的magnitude转移方案，使weight和activation的量化更容易。

$$\mathbf{s}_i = \max(|\mathbf{x}_i|) / \log_2(2 + \max(|\mathbf{x}_i|)); \quad \mathbf{x}_i = \mathbf{x}_i / \mathbf{s}_i \quad (1)$$

$$\mathbf{W}' = \mathrm{diag}(\mathbf{s})\mathbf{W}; \quad \mathbf{X}' = \mathbf{X}\mathrm{diag}(\mathbf{s})^{-1} \quad s.t. \quad \mathbf{X}'\mathbf{W}' = \mathbf{XW} \quad (2)$$

创新点

- 引入通道移位和缩放操作，以消除不对称并缩小异常通道。该算法能20分钟实现OPT-175B模型的离线量化。
- 在W8A8和W6A6设置下实现近乎无损量化



(a) Original distribution (b) Channel-wise shifting (c) Channel-wise scaling

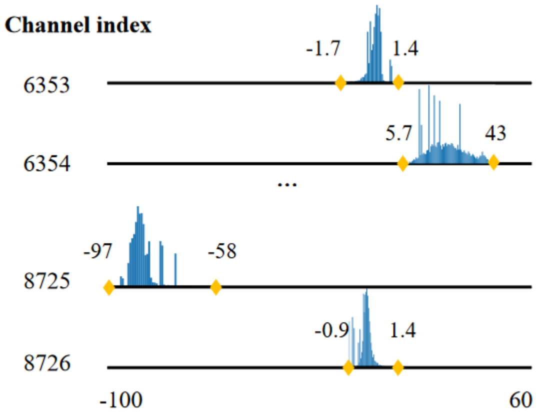
Outlier shifting and scaling

$$\widetilde{X'} = X - z, \qquad (1)$$

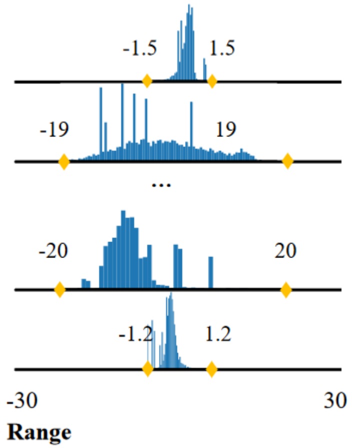
LLM中activation通道的异常值呈现不对称性，OPT-66B中第8725 channel范围在[-97, -58]，第6354 channel范围[5.7, 43]，呈现极端不对称性，仅使用min-max整个tensor的范围为[-97, 43]， 进行shifting后tensor范围[-20, 20]。

$$\widetilde{X} = (X - z) \oslash s. \qquad (2)$$

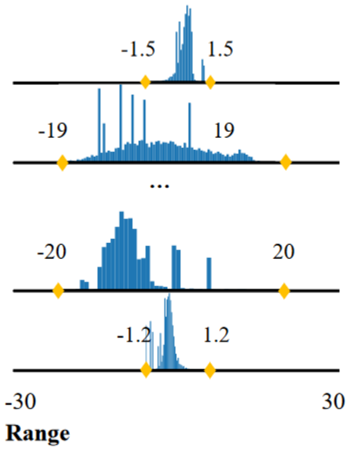
Scaling进一步缓解activation异常channel数据分布



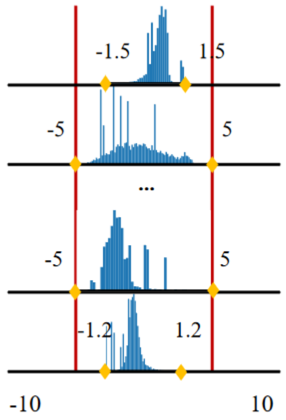
(a) Original distribution



(b) Channel-wise shifting



(b) Channel-wise shifting



(c) Channel-wise scaling

Unified migration pattern

$$Y = WX + b \longrightarrow \tilde{Y} = \tilde{X}\tilde{W} + \tilde{b} = \left[\frac{X - z}{s}\right][s \odot W] + (zW + b)$$

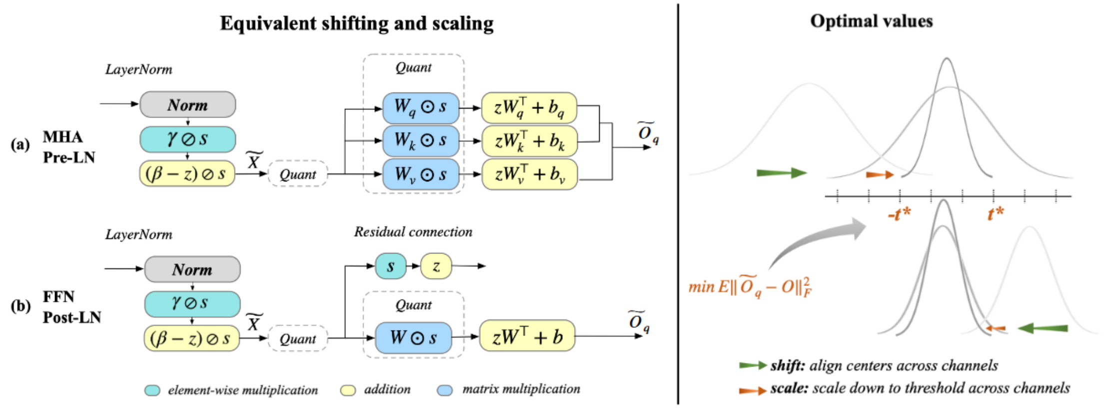


Figure 2: **Left**: We show the equivalent shifting and scaling operations by giving two representative examples: (a) for problematic output of Pre-LN (LayerNorm put inside residual connection) with Multi-Head Attention (MHA) structure; (b) for problematic output of Post-LN (LayerNorm put before residual connection) with Feed-Forward Network (FFN). **Right**: For optimal shifting and scaling values, the shifting vector can align the center of each channel to 0 and the scaling vector would shrink outliers into the outlier threshold t which is searched based on its left metric.

Algorithm 1: Outlier Suppression+

Input: Problematic output $\boldsymbol{X}$ of LayerNorm with parameters $\boldsymbol{\gamma}, \boldsymbol{\beta}$, subsequent module M with weight $\boldsymbol{W}$ and bias $\boldsymbol{b}$, grid search iteration K.

{1. Optimal shifting and scaling:}

$\boldsymbol{z} = \frac{\min(\boldsymbol{X}_{:,j}) + \max(\boldsymbol{X}_{:,j})}{2}$ ▷ Optimal shifting vector.

$loss^* = \mathrm{INF}$

for $k = 1$ *to* K **do**

 $t = \max(\boldsymbol{X} - \boldsymbol{z}) \cdot \frac{k}{K}$, ▷ Enumerate outlier threshold.

 $\boldsymbol{s}_j = \max(1.0, \frac{\max(\boldsymbol{X}_{:,j} - \boldsymbol{z}_j)}{t})$

 Calculate $loss_k$ based on Eq. (6), Eq. (7).

 if $loss^* > loss_k$ **then**

  $loss^* = loss_k$, $\boldsymbol{s}^* = \boldsymbol{s}$ ▷ Optimal scaling factors.

{2. Equivalent shifting and scaling:}

$\widetilde{\boldsymbol{\beta}} = (\boldsymbol{\beta} - \boldsymbol{z}) \oslash \boldsymbol{s}^*, \widetilde{\boldsymbol{\gamma}} = \boldsymbol{\gamma} \oslash \boldsymbol{s}^*_j$ ▷ Fuse $\boldsymbol{z}, \boldsymbol{s}^*$ into former operations.

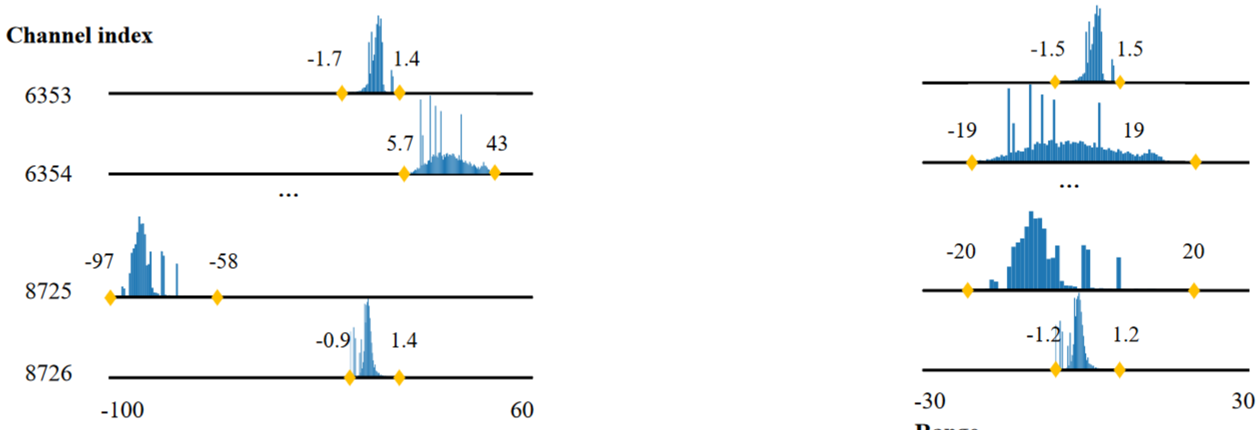
$\widetilde{\boldsymbol{b}} = \boldsymbol{z}\boldsymbol{W}^\top + \boldsymbol{b}, \widetilde{\boldsymbol{W}} = \boldsymbol{W} \odot \boldsymbol{s}^*$ ▷ Update following modules.

return Transformed LayerNorm and subsequent module;

How to choose shifting and scale?

$$z_j = \frac{\max(\boldsymbol{X}_{:,j}) + \min(\boldsymbol{X}_{:,j})}{2}. \qquad (5)$$

利用z_j实现channel数据以0为中心分布，保证数据分布的对称性。



(a) Original distribution

(b) Channel-wise shifting

$$s_j = \max(1.0, \frac{\max(\boldsymbol{X}_{:,j} - z_j)}{t}). \qquad (8)$$

$$\min_{\boldsymbol{s}} \mathbb{E}[\| \underbrace{Q((\boldsymbol{X}-\boldsymbol{z}) \oslash \boldsymbol{s}) Q(\boldsymbol{W} \odot \boldsymbol{s})^{\top} + \widetilde{\boldsymbol{b}}}_{\textit{output after scaling and quantization}} - \underbrace{(\boldsymbol{X}\boldsymbol{W}^{\top} + \boldsymbol{b})}_{\textit{original FP output}} \|_F^2], \qquad (6)$$

通过outlier threshold t，将对每个channel scale的选择，转化为对单一阈值变量t的选择。(t通过网格搜索实现)

小模型量化实验

Method	CoLA	MNLI	QNLI	SST-2	STS-B	Avg.
FP32	59.6	84.9	91.8	93.4	89.5	83.8
INT8*						
MinMax	52.3	81.3	89.0	91.1	86.2	79.5
OMSE	54.8	82.1	89.7	91.3	87.7	81.6
PEG	59.4	81.3	91.1	92.7	87.9	82.5
OS	60.3	83.9	90.2	**92.9**	88.2	83.0
Ours	**60.9**	**84.4**	**91.1**	92.7	**88.3**	**83.5**
INT6						
OMSE	35.4	73.7	84.7	86.3	85.8	73.5
Percentile	37.3	72.1	79.4	87.3	86.8	72.9
OS	54.4	81.8	89.8	91.9	88.7	81.2
Ours	**56.0**	**84.5**	**90.9**	**92.4**	**89.5**	**82.8**
INT4						
OMSE	4.7	38.5	52.2	50.3	0.2	41.1
Percentile	7.0	53.0	61.5	77.1	66.1	57.0
OS	28.5	57.9	72.5	80.4	67.8	62.7
Ours	**50.0**	**80.2**	**85.4**	**91.4**	**86.5**	**78.2**

Table 1: PTQ performance of BERT-base models. MNLI and STS-B report the combined score. **Avg.** indicates the averaged results of 8 tasks on GLUE benchmark (details in Appendix B). $*$ means per-tensor quantization for weight. OS indicates Outlier Suppression for short.

Method	CoLA (Matt.)	MNLI (acc m/mm)	MRPC (f1/acc)	QNLI (acc)	QQP (f1/acc)	RTE (acc)	SST-2 (acc)	STS-B (Pear./Spear.)	Avg.
FP32	63.3	86.7/85.9	91.6/88.0	92.2	88.1/91.1	74.0	93.5	90.3/90.1	84.9
INT8*									
MinMax	62.4	72.0/73.0	76.3/72.8	87.0	66.5/80.4	46.9	92.2	58.6/52.1	71.5
OMSE	59.9	82.7/83.5	87.8/83.8	89.0	79.2/86.2	47.3	92.0	83.9/83.3	78.1
Percentile	61.3	84.5/84.0	91.6/88.9	91.6	85.9/89.4	69.3	92.4	88.3/88.1	83.1
OS	**62.3**	85.1/84.5	90.1/86.0	91.1	87.0/90.3	**75.1**	92.4	88.7/88.4	83.9
Ours	62.2	**85.9/85.2**	**90.9/87.0**	**92.2**	**87.8/90.8**	71.8	**93.3**	**89.3/89.3**	**84.1**
INT6									
MinMax	5.6	32.0/32.0	50.2/46.1	50.2	0.0/63.2	49.5	53.0	5.0/4.8	38.1
OMSE	14.0	59.3/58.4	86.1/78.7	79.5	52.5/73.5	54.9	74.8	44.0/37.9	59.8
Percentile	16.4	63.5/63.8	82.0/77.2	87.0	44.8/70.7	49.8	81.7	65.7/67.8	62.8
OS	24.1	71.3/71.7	85.5/79.4	80.8	68.8/78.3	47.3	82.3	61.1/62.0	65.4
Ours	**60.9**	**86.3/85.4**	**91.8/88.2**	**92.0**	**87.7/90.8**	**71.5**	**93.7**	**86.7/85.6**	**83.7**

Table 7: PTQ performance of BERT-large models on GLUE benchmark. $*$ means per-tensor quantization for weight. OS indicates Outlier Suppression for short.

小模型量化实验

Method	CoLA (Matt.)	MNLI (acc m/mm)	MRPC (f1/acc)	QNLI (acc)	QQP (f1/acc)	RTE (acc)	SST-2 (acc)	STS-B (Pear./Spear.)	Avg.
FP32	59.6	84.9/84.8	91.4/87.8	91.8	87.8/90.9	72.6	93.4	89.7/89.3	83.8
INT8*									
MinMax	52.3	80.9/81.7	85.3/80.9	89.0	84.8/88.6	68.2	91.1	84.7/87.6	79.5
OMSE	54.8	81.9/82.2	89.7/86.0	89.7	86.1/89.5	72.2	91.3	87.2/88.2	81.6
PEG	59.4	81.3	88.5	91.1	**89.4**	69.3	92.7	87.9	82.5
OS	60.3	83.8/84.0	90.4/87.0	90.2	87.3/90.4	71.1	**92.9**	87.8/88.7	83.0
Ours	**60.9**	**84.4/84.4**	**90.6/87.2**	**91.1**	87.1/90.6	**73.3**	92.7	**87.7/88.9**	**83.5**
INT8									
MinMax	57.1	82.8/83.5	89.9/85.8	90.8	87.8/90.7	69.7	92.8	86.8/88.6	82.3
OMSE	57.2	84.0/84.3	90.1/85.8	91.1	87.6/90.5	72.2	92.2	87.9/88.7	82.9
Percentile	57.1	83.9/84.1	90.7/86.7	91.3	87.7/90.7	71.1	93.4	87.7/88.7	82.9
OS	**61.6**	84.4/84.5	**91.4/87.8**	91.5	**87.9/90.8**	**72.2**	**93.8**	89.2/89.0	**84.0**
Ours	60.3	**84.8/84.5**	90.5/87.0	**91.6**	87.5/90.8	71.5	93.6	**89.3/89.2**	83.6
INT6									
MinMax	17.7	32.5/32.5	0.7/31.9	65.2	40.9/69.0	48.0	82.0	59.8/60.3	47.1
OMSE	35.4	74.0/73.3	81.5/76.5	84.7	76.1/82.1	64.3	86.3	85.6/86.1	73.5
Percentile	37.3	72.4/71.7	85.1/79.9	79.4	72.6/80.2	61.7	87.3	86.4/87.3	72.9
OS	54.4	82.0/81.7	87.5/83.3	89.8	84.7/88.9	70.8	91.9	88.7/88.6	81.2
Ours	**56.0**	**84.6/84.4**	**90.0/86.3**	**90.9**	**87.0/90.5**	**71.8**	**92.4**	**89.6/89.4**	**82.8**
INT4									
MinMax	-6.6	32.6/32.7	0.0/31.6	50.6	53.8/36.8	47.7	50.9	-0.5/-0.5	29.5
OMSE	4.7	38.5/38.4	81.3/69.1	52.2	45.2/50.9	59.9	50.3	0.1/-0.4	41.1
Percentile	7.0	52.6/53.5	83.0/75.7	61.5	44.7/68.3	55.6	77.1	65.9/66.3	57.0
OS	28.5	57.5/58.3	83.9/75.7	72.5	45.4/70.8	56.7	80.4	67.8/67.9	62.7
Ours	**50.0**	**80.6/79.9**	**87.6/83.1**	**85.4**	**85.0/77.5**	**65.7**	**91.4**	**86.4/86.5**	**78.2**

Table 6: PTQ performance of BERT-base models on GLUE benchmark. $*$ means per-tensor quantization for weight. OS indicates Outlier Suppression for short.

两类大模型上的Zero-Shot实验对比

Name	Method	OPT-13B				OPT-30B				OPT-66B				OPT-175B			
		FP16	INT8*	INT8	INT6	FP16	INT8*	INT8	INT6	FP16	INT8*	INT8	INT6	FP16	INT8*	INT8	INT6
PIQA	LLM.int8()♣	75.8	-	75.8	-	77.6	-	77.3	-	78.7	-	78.7	-	79.7	-	79.6	-
	ZeroQuant♣		54.1	-	53.0		54.2	-	52.0		53.2	-	51.9		52.3	-	53.1
	SmoothQuant		76.0	-	73.5		77.2	-	66.7		78.3	-	52.0		**79.7**	-	52.6
	Ours		**76.4**	75.9	**75.8**		**77.4**	77.6	**77.4**		**78.7**	78.6	**77.5**		79.6	79.5	**80.0**
LAMBADA	LLM.int8()♣	68.6	-	68.4	-	71.5	-	71.4	-	73.9	-	73.8	-	74.7	-	74.6	
	ZeroQuant♣		0.0	-	0.0		0.0	-	0.0		0.0	-	0.0		0.0	-	0.0
	SmoothQuant		68.3	-	65.2		**71.0**	-	13.4		72.9	-	0.0		**74.6**	-	0.5
	Ours		**68.3**	68.4	**65.7**		70.8	70.8	**69.6**		**73.0**	73.4	**72.7**		74.5	74.5	**74.2**
HellaSwag	LLM.int8()♣	52.5	-	52.4	-	54.3	-	54.3	-	56.4	-	56.3	-	59.3	-	59.2	-
	ZeroQuant♣		26.5	-	25.8		26.4	-	25.7		26.1	-	25.7		25.4	-	25.6
	SmoothQuant		52.2	-	49.2		54.2	-	37.4		55.9	-	26.5		58.9	-	26.0
	Ours		**52.3**	52.5	**51.7**		**54.2**	54.2	**53.7**		**56.2**	56.3	**55.8**		**59.2**	59.3	**58.5**
Winogrande	LLM.int8()♣	65.1	-	64.8	-	68.5	-	68.1	-	68.9	-	68.5	-	72.5	-	72.3	-
	ZeroQuant♣		52.1	-	51.1		51.8	-	51.8		50.7	-	48.0		50.2	-	49.1
	SmoothQuant		64.9	-	60.3		**68.2**	-	55.0		68.3	-	52.1		71.2	-	49.1
	Ours		**65.0**	65.3	**64.0**		68.0	68.5	**68.9**		**69.0**	68.8	**69.4**		**72.5**	72.5	**71.7**
ARC (Challenge)	LLM.int8()♣	32.8	-	33.5	-	34.6	-	34.7	-	37.3	-	37.0	-	40.3	-	40.9	-
	ZeroQuant♣		19.3	-	20.7		19.8	-	20.6		20.8	-	20.4		21.8	-	20.6
	SmoothQuant		32.1	-	30.6		33.8	-	26.7		36.5	-	21.9		**40.5**	-	21.2
	Ours		**33.5**	33.3	**32.7**		**34.5**	34.7	**34.6**		**37.5**	37.2	**37.0**		40.3	39.9	**41.0**
ARC (Easy)	LLM.int8()♣	67.3	-	67.3	-	70.1	-	69.7	-	71.7	-	71.8	-	74.9	-	74.8	-
	ZeroQuant♣		27.5	-	25.0		30.5	-	25.0		29.7	-	26.0		24.0	-	25.6
	SmoothQuant		66.2	-	62.2		69.7	-	55.8		70.5	-	27.8		74.1	-	28.8
	Ours		**67.3**	66.8	**67.0**		**70.1**	70.0	**68.9**		**71.3**	71.8	**70.7**		**74.8**	74.7	**74.3**
COPA	LLM.int8()♣	86.0	-	86.0	-	82.0	-	82.0	-	86.0	-	87.0	-	88.0	-	89.0	-
	ZeroQuant♣		63.0	-	55.0		55.0	-	55.0		53.0	-	52.0		60.0	-	55.0
	SmoothQuant		85.0	-	82.0		83.0	-	75.0		84.0	-	55.0		88.0	-	55.0
	Ours		**85.0**	86.0	**85.0**		**83.0**	82.0	**84.0**		**85.0**	86.0	**84.0**		**88.0**	89.0	**91.0**
StoryCloze	LLM.int8()♣	76.1	-	76.3	-	77.0	-	77.1	-	77.5	-	77.7	-	79.5	-	79.3	-
	ZeroQuant♣		49.6	-	48.3		48.5	-	48.0		49.2	-	48.4		47.7	-	48.2
	SmoothQuant		**76.0**	-	73.5		76.9	-	61.4		77.3	-	48.8		79.1	-	49.8
	Ours		75.8	76.0	**75.4**		**77.0**	76.9	**76.6**		**77.3**	76.4	**76.6**		**79.2**	79.1	**78.1**
Avg.	Ours	65.5	65.5	65.5	64.7	67.0	66.9	66.8	66.7	68.8	68.5	68.6	68.0	71.1	71.0	71.1	71.1

Method	Bits	HellaSwag	LAMBADA	PIQA	Winogrande
BLOOM-176B	FP16	55.9	67.7	78.8	70.3
ZeroQuant♣	INT8*	54.8	67.8	76.0	**69.4**
SmoothQuant	INT8*	54.1	**69.2**	77.7	68.6
Ours	INT8*	**54.9**	68.0	**78.4**	69.1
ZeroQuant♣	INT6	30.5	7.5	61.2	52.0
SmoothQuant	INT6	52.1	60.2	76.7	67.6
Ours	INT6	**55.1**	**69.1**	**78.1**	**68.1**
BLOOMZ-176B	FP16	57.1	67.8	80.6	72.5
ZeroQuant♣	INT8*	56.3	67.6	79.1	70.9
SmoothQuant	INT8*	56.3	**68.7**	79.7	70.8
Ours	INT8*	**56.7**	68.5	**79.9**	**71.3**
ZeroQuant♣	INT6	28.2	1.4	54.0	49.6
SmoothQuant	INT6	55.0	65.2	**80.0**	69.9
Ours	INT6	**56.2**	**69.2**	79.9	**70.6**

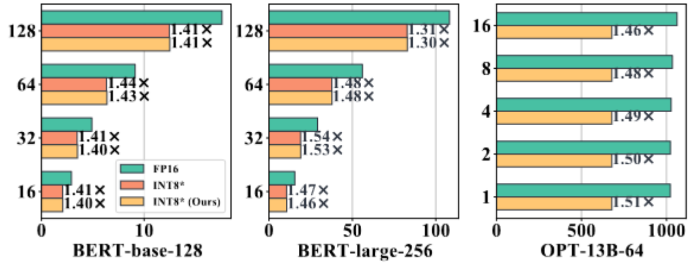
Table 5: Quantization results on 4 zero-shot tasks in terms of accuracy.

Ablation Study

Method	OPT-66B (INT6)		BERT (INT4)	
	PIQA	Winogrande	SST-2	MNLI
Ours	**77.5**	**69.4**	**91.4**	**80.2**
- shifting	76.5	66.5	89.3	77.7
- shifting - scaling	54.7	49.4	82.3	63.7

Table 4: Effect of scaling and shifting operations.

INT8量化下实际加速实验



量化内存减少实验

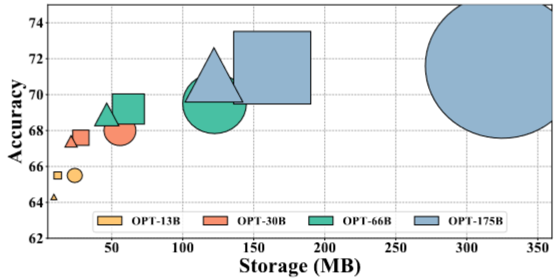
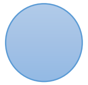


Figure 5: Averaged accuracy on PIQA, Winogrande, LAMBADA, and HellaSwag of OPTs with different storages. We draw circles, rectangles, and triangles to refer to FP16, the 8-bit and 6-bit models with quantized activation and weight.


Fp16 model


INT8 model


INT6 model

THNAKS FOR LISTENING