QA-LoRA: Quantization-Aware Low-Rank Adaptation of Large Language Models

Yuhui Xu Lingxi Xie Xiaotao Gu Xin Chen Heng Chang
Hengheng Zhang Zhensu Chen Xiaopeng Zhang Qi Tian

Huawei Inc.

Introduction

LLMs have shown unprecedented performance across a wide range of language understanding tasks and served as the foundation of state-of-the-art chat systems. The diversity of real-world applications calls for a pipeline in which LLMs can be fine-tuned to fit different scenarios and quantized to be deployed onto edge devices (*e.g.*, mobile phones), and the key issue is to get rid of the heavy computational burden brought by the large number of parameters of LLMs.

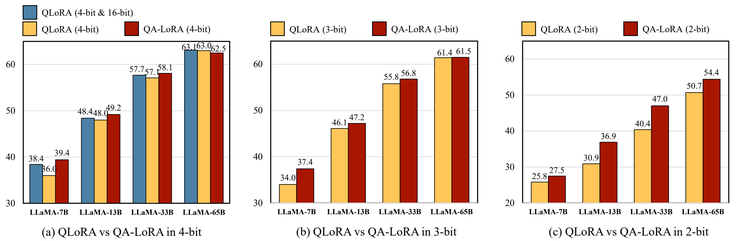

Figure 1. The comparison of 5-shot MMLU accuracy (%) with different quantization bit widths based on the LLaMA model family

There are two lines of research for this purpose:

- **The first one** is parameter-efficient fine-tuning (PEFT) which introduced a small number of learnable parameters while keeping most pre-trained parameters unchanged. Among them, low-rank adaptation (LoRA), a popular PEFT algorithm, proposed to fine-tune low-rank matrices to complement the pre-trained weights. Despite the comparable performance to full-parameter fine-tuning, the memory usage of LoRA is still large, especially when the base LLM is large (*e.g.*, LLaMA-65B).
- **The second one** studies parameter quantization where the trained weights are quantized into low-bit integers or floating point numbers. Although these methods can alleviate the computational burden, they often report unsatisfying accuracy especially when the quantization bit width is low.

Contributions

it is an important topic to integrate PEFT with quantization. A naive solution is to perform post-training quantization (PTQ) after PEFT, but it reports unsatisfying accuracy especially when the quantization bit width is low. QA-LoRA addresses the issue by introducing group-wise operators to increase the number of parameters for low-bit quantization (each group is quantized individually) and decrease that of LoRA (each group shares the adaptation parameters). QA-LoRA enjoys two-fold benefits:

- An efficient fine-tuning stage thanks to the LLM's weights being quantized into low-bit integers;
- A lightweight, fine-tuned model without the need for PTQ which often incurs loss of accuracy

The framework of QA-LoRA

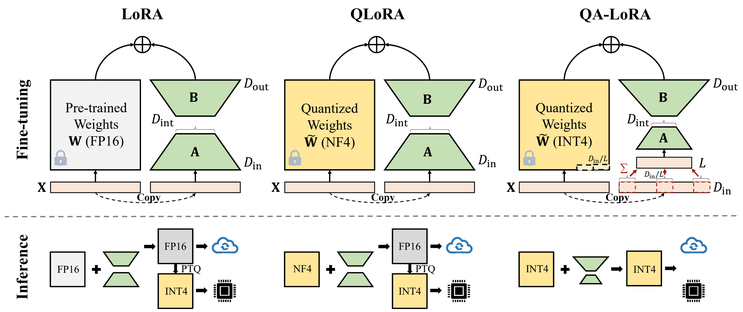

Figure 2. An illustration of the goal of QA-LoRA. Compared to prior adaptation methods, LoRA and QLoRA, our approach is computationally efficient in both the fine-tuning and inference stages. More importantly, it does not suffer an accuracy loss because post-training quantization is not required. We display INT4 quantization in the figure, but QA-LoRA is generalized to INT3 and INT2.

Methodology

LoRA:

The key idea of LoRA is to introduce a pair of matrices, $\mathbf{A}$ and $\mathbf{B}$, to supplement $\mathbf{W}$. $\mathbf{A}$ and $\mathbf{B}$ have sizes of $D_{\text{in}} \times D_{\text{int}}$ and $D_{\text{int}} \times D_{\text{out}}$, respectively. The intermediate dimensionality is often small (*i.e.*, $D_{\text{int}} \ll \min\{D_{\text{in}}, D_{\text{out}}\}$), making $\mathbf{AB}$ a low-rank matrix compared to $\mathbf{W}$.

- Fine-tuning, $\mathbf{y} = \mathbf{W}^\top \mathbf{x} + s \cdot (\mathbf{AB})^\top \mathbf{x}$
- Inference, $\mathbf{y} = (\mathbf{W} + s \cdot \mathbf{AB})^\top \mathbf{x}$

Quantization:

Mathematically, given the bit width N and a pre-trained weight matrix $\mathbf{W}$, we compute the minimum and maximum values across all elements of $\mathbf{W}$, denoted as $\min(\mathbf{W})$ and $\max(\mathbf{W})$, respectively. Then, $\mathbf{W}$ is quantized into $\tilde{\mathbf{W}}$ by computing

$$\tilde{\mathbf{W}} = \alpha \cdot \hat{\mathbf{W}} + \beta \doteq \alpha \cdot \left\lfloor \frac{\mathbf{W} - \beta}{\alpha} \right\rceil + \beta,$$

where $\alpha = (\max(\mathbf{W}) - \min(\mathbf{W}))/(2^N - 1)$ and $\beta = \min(\mathbf{W})$ are called the scaling and zero factors, respectively.

QA-LoRA:

Our objective is efficient adaptation and deployment. The key to achieving the second goal lies in that $\tilde{\mathbf{W}}$ (*i.e.*, the quantized $\mathbf{W}$) and $s \cdot \mathbf{AB}$ can be merged without using high-precision numbers (*e.g.*, FP16). QA-LoRA ntroduce group-wise operations for both quantization and the input features of LoRA, such that LoRA parameters can be merged.

$$\beta' = \beta + s \cdot \mathbf{AB},$$

Experiments

Table 1. 0-shot and 5-shot accuracy (%) on the Massive Multitask Language Understanding (MMLU) dataset

Method	Dataset	#Bits	MMLU (0-shot) Hums.	STEM	Social	Other	Avg.	MMLU (5-shot) Hums.	STEM	Social	Other	Avg.
LLaMA-7B	–	16	32.4	26.6	31.4	37.2	32.1	33.3	29.8	37.8	38.0	34.6
QLoRA	*Alpaca*	*4+16*	*38.1*	*31.1*	*41.6*	*46.9*	*39.4*	*36.1*	*31.9*	*42.0*	*44.5*	*38.4*
QLoRA w/ GPTQ	Alpaca	4	35.7	30.9	38.0	44.0	37.1	33.8	31.3	37.4	42.2	36.0
PEQA	Alpaca	4	–	–	–	–	–	34.9	28.9	37.5	40.1	34.8
QA-LoRA	Alpaca	4	36.9	31.4	40.3	44.9	**38.3**	36.6	32.4	44.8	44.9	**39.4**
QLoRA w/ GPTQ	Alpaca	3	31.5	28.9	31.8	36.8	32.2	31.6	30.1	35.6	39.8	34.0
QA-LoRA	Alpaca	3	36.0	34.1	42.0	42.3	**38.3**	35.6	30.5	41.5	42.7	**37.4**
QLoRA w/ GPTQ	Alpaca	2	24.1	22.1	22.5	23.7	23.2	23.4	26.2	26.4	28.4	25.8
QA-LoRA	Alpaca	2	26.4	25.5	25.6	28.7	**26.5**	27.3	26.1	26.1	30.3	**27.5**
QLoRA	*FLAN v2*	*4+16*	*40.9*	*32.5*	*47.8*	*49.5*	*42.6*	*41.4*	*35.0*	*49.8*	*52.0*	*44.3*
QLoRA w/ GPTQ	FLAN v2	4	39.7	32.5	46.4	48.1	41.6	36.5	33.7	46.9	50.3	41.4
QA-LoRA	FLAN v2	4	44.0	35.3	52.3	52.6	**45.9**	43.9	38.0	54.3	53.0	**47.0**
QLoRA w/ GPTQ	FLAN v2	3	36.7	30.2	38.4	40.1	36.5	32.2	31.7	42.7	42.8	36.9
QA-LoRA	FLAN v2	3	41.4	35.1	52.0	50.2	**44.4**	41.3	36.0	52.8	50.2	**44.7**
QLoRA w/ GPTQ	FLAN v2	2	24.1	22.5	22.3	23.8	23.3	23.9	25.3	26.2	25.3	25.0
QA-LoRA	FLAN v2	2	34.1	30.0	37.2	39.8	**35.2**	31.8	38.1	34.5	38.5	**33.2**
LLaMA-13B	–	16	40.6	36.7	48.9	48.0	43.3	44.0	35.9	53.2	52.9	46.3
QLoRA	*Alpaca*	*4+16*	*45.2*	*38.3*	*55.0*	*54.6*	*48.1*	*46.0*	*37.3*	*55.8*	*55.1*	*48.4*
QLoRA w/ GPTQ	Alpaca	4	44.7	38.0	54.4	54.0	47.6	45.4	37.4	55.7	54.3	48.0
PEQA	Alpaca	4	–	–	–	–	–	43.0	37.7	53.6	49.0	45.0
QA-LoRA	Alpaca	4	44.3	38.0	55.1	55.5	**47.9**	48.4	38.3	54.9	55.2	**49.2**
QLoRA w/ GPTQ	Alpaca	3	43.5	36.2	52.3	52.6	45.9	43.6	36.1	53.0	52.7	46.1
QA-LoRA	Alpaca	3	43.9	37.3	53.1	54.3	**46.9**	44.3	38.8	53.4	53.8	**47.3**
QLoRA w/ GPTQ	Alpaca	2	27.7	27.6	31.8	29.7	29.0	29.0	27.1	33.4	34.8	30.9
QA-LoRA	Alpaca	2	35.7	33.3	40.9	42.0	**37.8**	35.6	30.6	39.9	41.7	**36.9**
QLoRA	*FLAN v2*	*4+16*	*48.0*	*39.2*	*58.2*	*56.7*	*50.3*	*49.9*	*40.1*	*60.2*	*57.9*	*51.9*
QLoRA w/ GPTQ	FLAN v2	4	47.6	39.6	57.6	56.0	50.0	49.4	40.9	59.7	57.6	51.7
QA-LoRA	FLAN v2	4	47.7	41.4	59.6	57.2	**51.1**	50.0	41.5	60.5	58.4	**52.4**
QLoRA w/ GPTQ	FLAN v2	3	46.6	37.9	55.9	55.7	48.9	46.5	38.2	57.2	56.1	49.3
QA-LoRA	FLAN v2	3	47.4	39.4	57.7	56.0	**49.9**	49.3	40.0	60.0	57.5	**51.5**
QLoRA w/ GPTQ	FLAN v2	2	36.2	30.3	40.8	44.1	37.8	36.6	32.0	43.8	44.2	38.9
QA-LoRA	FLAN v2	2	40.8	36.4	39.3	50.1	**43.9**	40.9	36.1	50.7	46.7	**44.1**
LLaMA-33B	–	16	51.0	42.7	63.3	60.4	54.1	56.2	45.9	67.1	63.9	58.2
QLoRA	*Alpaca*	*4+16*	*52.2*	*44.9*	*64.3*	*61.8*	*55.5*	*55.4*	*46.0*	*66.4*	*63.6*	*57.7*
QLoRA w/ GPTQ	Alpaca	4	51.7	44.7	63.4	61.0	54.9	53.9	46.6	66.3	62.9	57.1
QA-LoRA	Alpaca	4	51.6	44.9	65.0	61.8	**55.4**	55.8	46.4	67.0	64.0	**58.1**
QLoRA w/ GPTQ	Alpaca	3	49.5	43.3	63.1	61.0	53.8	53.3	45.0	64.1	61.4	55.8
QA-LoRA	Alpaca	3	50.6	44.6	64.0	61.2	**54.7**	54.3	45.8	65.2	62.6	**56.8**
QLoRA w/ GPTQ	Alpaca	2	32.0	31.6	35.8	32.8	32.9	37.5	34.9	45.3	44.9	40.4
QA-LoRA	Alpaca	2	38.4	38.2	50.7	49.7	**43.6**	44.2	38.8	53.9	52.3	**47.0**
QLoRA	*FLAN v2*	*4+16*	*56.3*	*46.5*	*68.6*	*64.6*	*58.8*	*57.2*	*48.6*	*69.8*	*65.2*	*60.0*
QLoRA w/ GPTQ	FLAN v2	4	54.9	46.4	68.2	63.6	58.0	57.4	48.6	69.2	64.9	59.8
QA-LoRA	FLAN v2	4	54.2	47.0	69.7	65.5	**58.7**	57.9	48.8	71.0	65.5	**60.6**
QLoRA w/ GPTQ	FLAN v2	3	54.0	44.3	65.8	62.7	56.5	55.7	47.4	67.9	64.0	58.5
QA-LoRA	FLAN v2	3	53.1	45.0	66.9	63.0	**56.7**	56.8	46.9	68.9	63.7	**58.9**
QLoRA w/ GPTQ	FLAN v2	2	37.9	35.0	47.6	42.9	40.6	42.8	37.0	54.3	51.5	46.1
QA-LoRA	FLAN v2	2	49.4	40.4	59.8	56.5	**51.4**	49.6	42.7	60.7	57.8	**52.4**
LLaMA-65B	–	16	56.4	45.2	68.0	64.1	58.3	61.4	51.9	73.6	67.6	63.4
QLoRA	*Alpaca*	*4+16*	*55.5*	*49.3*	*70.4*	*66.9*	*60.1*	*60.3*	*52.7*	*72.9*	*67.4*	*63.1*
QLoRA w/ GPTQ	Alpaca	4	54.8	48.9	69.8	66.1	59.4	60.4	52.5	73.0	67.2	**63.0**
QA-LoRA	Alpaca	4	57.1	48.2	70.7	64.9	**60.0**	60.8	50.5	72.5	66.7	62.5
QLoRA w/ GPTQ	Alpaca	3	57.4	47.9	67.2	65.1	59.3	59.6	50.0	70.6	66.1	61.4
QA-LoRA	Alpaca	3	57.6	48.4	69.3	65.4	**60.0**	59.3	49.6	71.9	66.0	**61.5**
QLoRA w/ GPTQ	Alpaca	2	43.9	38.0	42.6	51.1	46.2	47.3	40.8	58.9	57.0	50.7
QA-LoRA	Alpaca	2	48.6	42.5	60.7	58.6	**52.2**	51.3	43.4	63.4	60.7	**54.4**
QLoRA	*FLAN v2*	*4+16*	*58.8*	*52.5*	*74.0*	*67.4*	*62.8*	*59.8*	*52.9*	*75.0*	*69.6*	*63.9*
[0.8pt/1pt] QLoRA w/ GPTQ	FLAN v2	4	57.8	51.9	73.5	67.8	62.3	59.2	52.5	75.0	69.3	**63.5**
QA-LoRA	FLAN v2	4	64.1	52.6	74.8	69.1	**65.1**	57.6	51.1	73.9	67.4	62.1
QLoRA w/ GPTQ	FLAN v2	3	58.5	50.2	71.5	66.9	**61.5**	59.9	51.7	73.4	67.9	63.0
QA-LoRA	FLAN v2	3	57.5	49.5	72.4	66.9	61.2	61.7	51.1	73.8	68.4	**63.6**
QLoRA w/ GPTQ	FLAN v2	2	47.9	43.1	60.1	56.0	51.4	52.6	43.8	62.8	58.5	54.3
QA-LoRA	FLAN v2	2	55.9	44.6	65.6	63.4	**57.1**	55.5	46.8	67.3	63.2	**58.0**