DL-QAT: Weight-Decomposed Low-Rank Quantization-Aware Training for Large Language Models

Wenjing Ke, Zhe Li, Dong Li, Lu Tian, Emad Barsoum
Advanced Micro Devices, Inc., Beijing, China

Abstract

Improving the efficiency of inference in Large Language Models (LLMs) is a critical area of research. Post-training Quantization (PTQ) is a popular technique, but it often faces challenges at low-bit levels, particularly in downstream tasks. Quantization-aware Training (QAT) can alleviate this problem, but it requires significantly more computational resources. To tackle this, we introduced Weight-**D**ecomposed **L**ow-Rank **Q**uantization-**A**ware **T**raining (DL-QAT), which merges the advantages of QAT while training only less than 1% of the total parameters. Specifically, we introduce a group-specific quantization magnitude to adjust the overall scale of each quantization group. Within each quantization group, we use LoRA matrices to update the weight size and direction in the quantization space. We validated the effectiveness of our method on the LLaMA and LLaMA2 model families. The results show significant improvements over our baseline method across different quantization granularities. For instance, for LLaMA-7B, our approach outperforms the previous state-of-the-art method by 4.2% in MMLU on 3-bit LLaMA-7B model. Additionally, our quantization results on pre-trained models also surpass previous QAT methods, demonstrating the superior performance and efficiency of our approach.

Method

In large language models (LLMs), LoRA (Low-Rank Adaptation) refines the model by introducing two low-rank matrices A and B. The weight matrix W can be presented as:

$$W = W_0 + \alpha BA$$

where W_0 represents the pretrained weight matrix that remains frozen during training, and α is a scaling factor that adjusts the influence of the low-rank adaptation.

For a given bit level n, the asymmetric weight quantization and dequantization processes can be described by a specific formula:

$$\widetilde{w} = clip(\left\lfloor \frac{W - b}{s} \right\rfloor, -2^{n-1}, 2^{n-1} - 1)$$

$$W_q = s * \widetilde{w} + b$$

where $\widetilde{w}$ represents the quantized value. The scale s determines the step size between quantization levels, and b is the offset applied to the weight before scaling.

In DL-QAT, rather than updating the weight matrix W directly, the updates are applied to the LoRA matrices A and B. As a result, the quantization and de-quantization formula is modified accordingly:

$$\widetilde{w'} = clip(\left\lfloor \frac{W_0 + \alpha BA - b}{s} \right\rfloor, -2^{n-1}, 2^{n-1} - 1)$$

$$W_q' = s * \widetilde{w'} + b$$

We separate the joint training of LoRA and quantization into two parts: (1) group-specific magnitude training; (2) weight fine-tuning in the pre-defined quantization space. The quantization process is thus reformulated as follows:

$$\begin{aligned} \mathrm{W_q} &= m * \mathrm{W_q'} \\ &= m * (W_0 + \alpha BA)_q \\ &= m * (s * \left(\widetilde{W_0 + \alpha BA}\right) + b) \end{aligned}$$

Here, m represents a newly introduced hyper-parameter denoting the group-specific magnitude.

During the initial training phase, the scale factors s and the biases b are trained to ensure that the quantization updates commence from a well-established quantization space. Then adjust the magnitude term m to set the scale for each quantization group. At last the A and B matrices are fine-tuned, permitting updates to the quantized weights within the established quantization space.

Experiments

LLaMA	Method	Bits	MMLU 0-Shot	MMLU 5-Shot	Common Sense Zero-Shot ARC_C	ARC_E	BoolQ	HellaSwag	OBQA	PIQA	Winogrande	Avg
1-7B	-	16	32.1	34.6	38.2	67.3	72.9	56.3	28.4	78.2	67.1	58.3
	QA-LoRA*	4	37.9	38.5	44.0	71.6	75.9	57.1	30.8	78.9	67.2	60.8
	Ours	4	**40.5**	**39.9**	**45.0**	**75.5**	**79.8**	**57.9**	**36.2**	**78.9**	**70.2**	**63.4**
	QA-LoRA*	3	32.2	32.9	**41.7**	71.6	76.9	54.6	28.0	77.6	64.9	59.3
	Ours	3	**36.4**	**33.9**	41.0	**73.4**	**78.2**	**55.3**	**34.2**	**78.2**	**67.5**	**61.1**
2-7B	-	16	40.7	45.5	39.9	69.3	71.1	56.7	31.8	78.3	67.1	59.2
	QA-LoRA*	4	42.5	44.8	42.7	71.9	77.6	56.9	32.6	**79.2**	68.3	61.3
	Ours	4	**44.6**	**45.0**	**47.2**	**77.8**	**79.3**	**58.1**	**35.6**	78.5	**68.5**	**63.6**
	QA-LoRA*	3	37.9	37.9	38.1	66.6	75.0	54.0	32.0	76.0	66.5	58.3
	Ours	3	**40.5**	**39.4**	**41.2**	**74.4**	**78.0**	**54.7**	**32.2**	**77.5**	**68.8**	**60.9**

Table 1: Results of weight-only group-wise quantization with group_size=128 on LLaMA-7B and LLaMA2-7B.

LLaMA	Method	W-A-KV	Wikitext2 ppl (↓)	Common Sense Zero-Shot ARC_C	ARC_E	BoolQ	HellaSwag	PIQA	Winogrande	Avg
1-7B	-	16-16-16	5.68	48.0	73.0	76.8	76.1	79.3	70.0	70.5
	QA-LoRA*	3-16-16	16.5	38.4	51.5	64.3	64.5	73.7	60.9	58.9
	Ours	3-16-16	**9.2**	**40.1**	**61.8**	**71.2**	**67.2**	**75.9**	**64.0**	**63.4**
	QA-LoRA*	4-16-16	11.1	42.4	58.0	73.7	70.5	77.3	66.1	64.7
	LLM-QAT	4-16-16	-	**45.0**	**70.0**	75.5	74.0	**78.3**	**69.0**	68.6
	Ours	4-16-16	**6.7**	44.4	68.5	**78.5**	**74.4**	78.1	68.5	**68.7**
	SmoothQuant	4-8-8	-	42.8	67.4	71.0	67.8	77.6	66.0	65.2
	LLM-QAT	4-8-8	-	45.6	70.2	74.6	73.5	77.5	67.7	68.2
	Ours	4-8-8	**6.7**	**46.2**	**71.3**	**78.1**	**73.6**	**78.5**	**68.4**	**69.4**
1-13B	-	16-16-16	5.09	52.6	74.5	78.1	79.2	80.0	73.6	73.0
	SmoothQuant	4-8-8	-	43.3	67.4	72.5	74.3	77.1	69.5	67.4
	LLM-QAT	4-8-8	-	**51.9**	73.6	77.5	73.6	79.1	**70.6**	71.6
	Ours	4-8-8	**5.9**	48.8	**74.8**	**80.5**	**77.1**	**80.4**	70.3	**72.0**
2-7B	-	16-16-16	5.47	46.3	74.6	77.7	76.0	79.1	69.1	70.5
	QA-LoRA*	3-16-16	13.7	**36.3**	48.2	70.3	**66.3**	74.4	**63.9**	59.9
	Ours	3-16-16	**9.4**	35.9	**58.9**	**71.1**	63.6	**74.8**	60.2	**63.7**
	QA-LoRA*	4-16-16	9.5	41.3	55.1	68.8	71.9	77.3	68.2	63.8
	Ours	4-16-16	**6.3**	**44.6**	**71.0**	**78.5**	**74.6**	**78.2**	**68.8**	**69.3**
2-13B	-	16-16-16	4.88	49.0	77.4	80.6	79.4	80.5	72.2	73.2
	Ours	4-8-8	5.63	49.6	75.5	81.9	78.1	80.1	70.3	72.6

Table 2: Results of channel-wise quantization on LLaMA-7B/13B models.

LLaMA	Quant config	Trainable Params (M) s, b	Trainable Params (M) m, A, B	GPU Memory (G)	Training speed (s/iter)
7B	Weight-only, g128	50	71	32.5	3.33
	Weight-only, per-channel	1	41	31.8	3.24
	Quant W/A/KV, per-channel	1	41	33.1	3.91
13B	Weight-only, g128	99	162	60.4	6.26
	Weight-only, per-channel	2	65	58.7	6.09
	Quant W/A/KV, per-channel	2	65	62.8	7.04

Table 3: Training parameter count, GPU memory usage, and training speed for LLaMA-7B/13B under different quantization configurations with a per-GPU batch size of 16. The experiments were conducted on an AMD MI250 with 64GB of GPU memory.