On-the-Fly Adaptation to Quantization: Configuration-Aware LoRA for Efficient Fine-Tuning of Quantized LLMs

Rongguang Ye [1] Ming Tang [1] Edith C. H. Ngai [2]
[1]Southern University of Science and Technology, [2]The University of Hong Kong

Motivation

- **Performance Gap:** a single shared LoRA adapter cannot achieve consistently optimal performance across different quantization configurations.
- **Efficiency Issue:** Training separate LoRA adapters for multiple bit-widths is time-consuming.

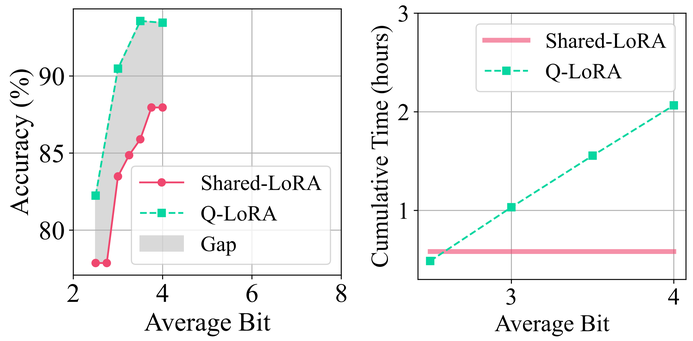

Figure 1. Accuracy gap (left) and cumulative fine-tuning time (right) on SST-2 using RoBERTa-Large.

Method

1. **Goal:** Effectively adjust LoRA adapters to compensate quantization errors under different layer-wise configurations.
2. **Configuration Representation:** Encode each layer's quantization parameters (bit-widths, buckets, layer type) into a vector embedding.
3. **Configuration-Aware Model:** Learn a mapping from layer embeddings to low-rank LoRA parameters.
4. **Pareto Guidance:** Select diverse, high-performing configurations via Pareto-based search.

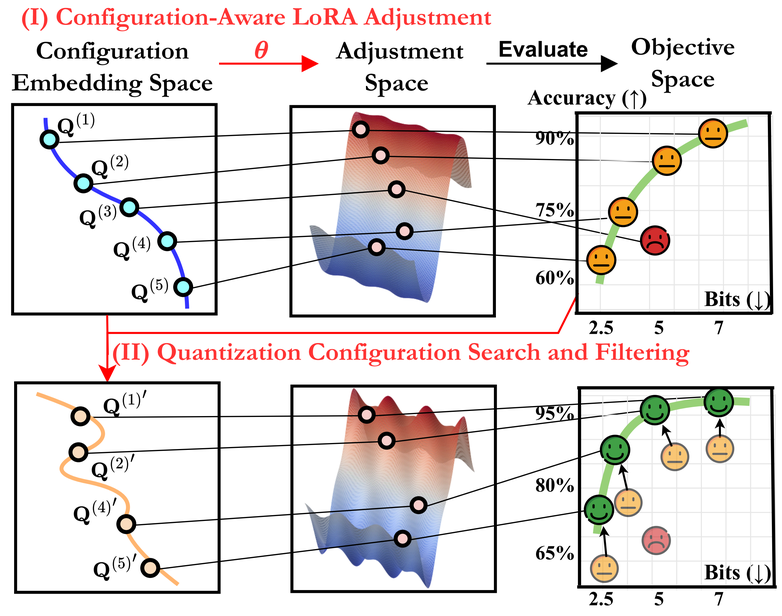

Figure 2. CoA-LoRA workflow: optimizing both the quantization configurations and the configuration-aware model to achieve maximum accuracy at any given bit-width.

Experiments

- **Robust Across Quantizations (Fig. 3 & Table 1):** CoA-LoRA maintains strong and stable accuracy under both NF4 quantization and integer mixed-precision.
- **Top Performance Across Models (Fig. 4):** CoA-LoRA achieves leading accuracy on LLMs ranging from 1.5B to 7B parameters, demonstrating effectiveness across different model sizes.
- **Single-Run Scalability:** CoA-LoRA adjusts LoRA adapters across different bit-widths with just one training run, avoiding repeated fine-tuning for each configuration.

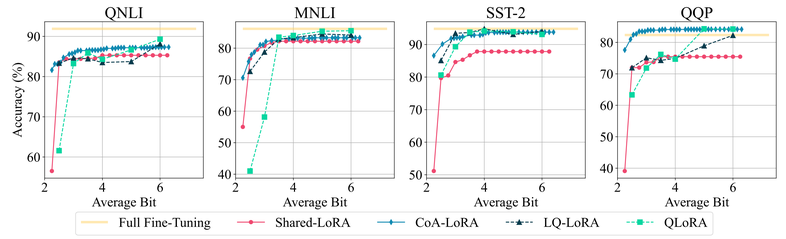

Figure 3. Performance comparison across average bit-widths.

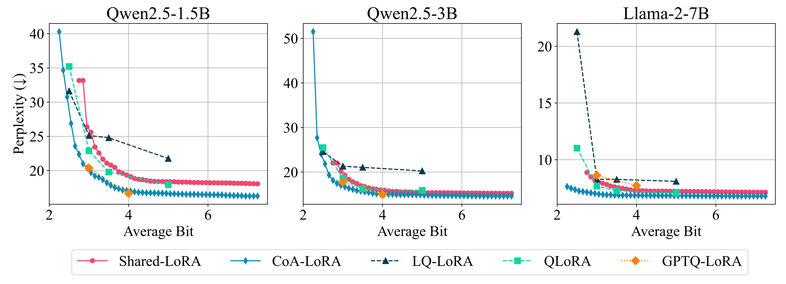

Figure 4. Perplexity results on causal LLMs of varying sizes.

Table 1. Accuracy comparison under integer mixed-precision quantization (per-layer choices: `int2`, `int3`, `int4`, `int8`). Columns correspond to methods, and values within each cell are reported in the order of tasks: QNLI / MNLI / SST-2.

Avg. Bit	QLoRA	LQ-LoRA	Shared-LoRA	CoA-LoRA
3	0.8537 / 0.8153 / 0.8291	0.8552 / 0.7397 / **0.9208**	0.7772 / 0.6618 / 0.8199	**0.8616** / 0.8099 / 0.8589
4	0.8828 / 0.8491 / 0.9116	0.8762 / 0.7913 / 0.9185	0.8762 / 0.7632 / 0.8486	**0.8841** / 0.8441 / **0.9254**
5	0.8859 / 0.8554 / 0.9105	0.8761 / 0.8481 / **0.9369**	0.8777 / 0.7674 / 0.8383	**0.8863** / 0.8533 / 0.9346
6	0.8833 / 0.8466 / 0.9174	0.8744 / 0.8484 / **0.9323**	0.8790 / 0.7704 / 0.8440	**0.8925** / **0.8558** / **0.9323**

Conclusions

- **Real-Time Adaptation:** Our proposed method enables on-the-fly adjustment of low-rank adapters to arbitrary quantization configurations without extra fine-tuning.
- **Robust and Generalizable:** CoA-LoRA maintains stable performance across tasks and generalizes to unseen configurations, supporting efficient LLM deployment on heterogeneous devices.

yerg2023@mail.sustech.edu.c