MLoRQ: Bridging Low-Rank and Quantization for Transformer Compression

Ofir Gordon, Ariel Lapid, Elad Cohen, Yarden Yagil, Arnon Netzer, Hai Habi

BMVC 2025

Overview

MLoRQ–Mixed Low-Rank and Quantization, is a **post-training compression framework** targeted mainly for Transformers leverages their matrix-heavy structure. MLoRQ *jointly* applies low-rank approximation and mixed-precision quantization under a unified optimization, achieving a superior balance of efficiency and accuracy.

Our method delivers **state-of-the-art compression**, reducing models to less than 12.5% of their size while improving accuracy by up to 15% over leading baselines. It is simple to integrate with existing quantization methods and has been validated across vision and language tasks.

Problem Statement

Let $\boldsymbol{f}$ be a pre-trained neural network, $\mathcal{S} \triangleq \left(\widetilde{\boldsymbol{W}}_1, \dots, \widetilde{\boldsymbol{W}}_L\right)$, where $\widetilde{\boldsymbol{W}}_\ell \in \mathcal{W}_\ell$, denotes the compressed weight assignment, $\Delta \boldsymbol{w}$ is the vector representing the difference between the float and compressed weights. Given the constraint of ψ_{WMS} for the weights memory size, we give the following problem formulation:

$$\min_{\mathcal{S}} \mathcal{L}\left(\boldsymbol{w} + \Delta\boldsymbol{w}\left(\mathcal{S}\right)\right),$$
$$\text{s.t.} : \text{WeightsMemory}\left(\mathcal{S}\right) \le \psi_{WMS}.$$

Unified Representation

- Let $\boldsymbol{W}_\ell = \boldsymbol{A}_\ell \boldsymbol{B}_\ell$ be a decomposition of the layer's weights.
- r_ℓ is a given rank and $b_\ell^{\boldsymbol{A}}, b_\ell^{\boldsymbol{B}} \in \mathcal{B}$ are bit-widths to quantize $= \boldsymbol{A}_\ell$ and $\boldsymbol{B}_\ell$.

The sets of compressed weights options are:

$$\mathcal{W}_\ell^{Q} \triangleq \left\{ Q\left(\boldsymbol{W}_\ell, \phi_{\boldsymbol{W}_\ell}\left(b_{\boldsymbol{W}_\ell}\right), b_{\boldsymbol{W}_\ell}\right) \mid b_{\boldsymbol{W}_\ell} \in \mathcal{B} \right\}$$
$$\mathcal{W}_\ell^{LQ} \triangleq \left\{ \widetilde{\boldsymbol{A}}_\ell^{(r)}\left(b_\ell^{\boldsymbol{A}}\right) \widetilde{\boldsymbol{B}}_\ell^{(r)}\left(b_\ell^{\boldsymbol{B}}\right) \mid b_\ell^{\boldsymbol{A}}, b_\ell^{\boldsymbol{B}} \in \mathcal{B}, r_\ell \in \mathcal{R}_\ell \right\},$$

where $\widetilde{\boldsymbol{A}}_\ell^{(r)}$ and $\widetilde{\boldsymbol{B}}_\ell^{(r)}$ are the compressed matrices.

Joint vs. Quantiztion Only

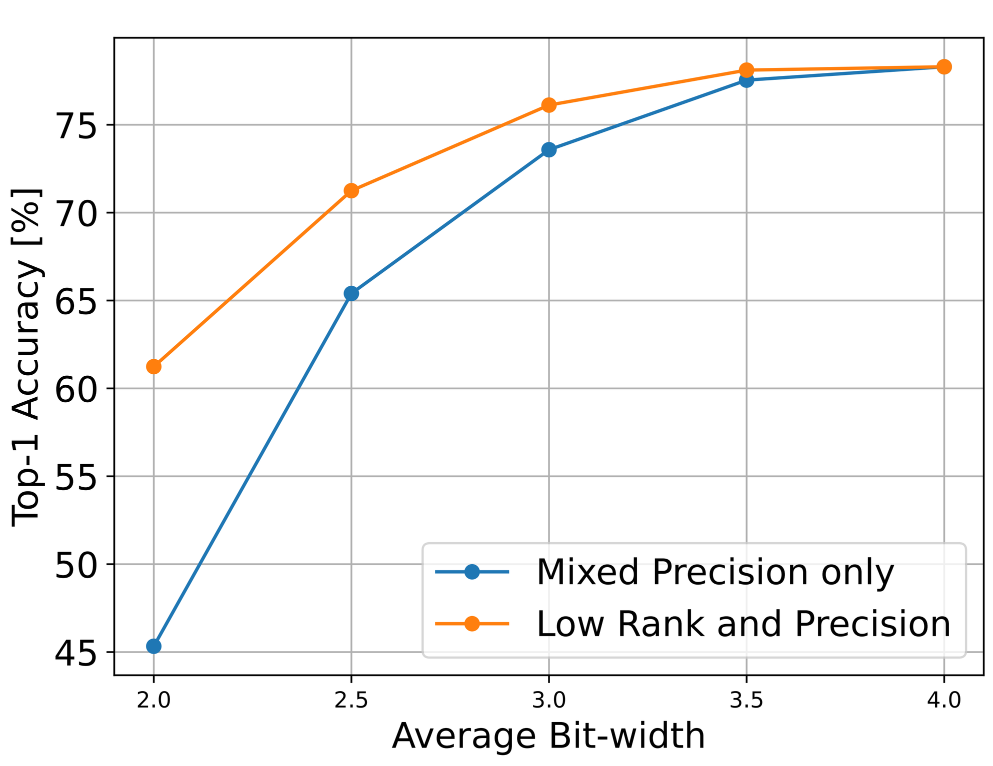
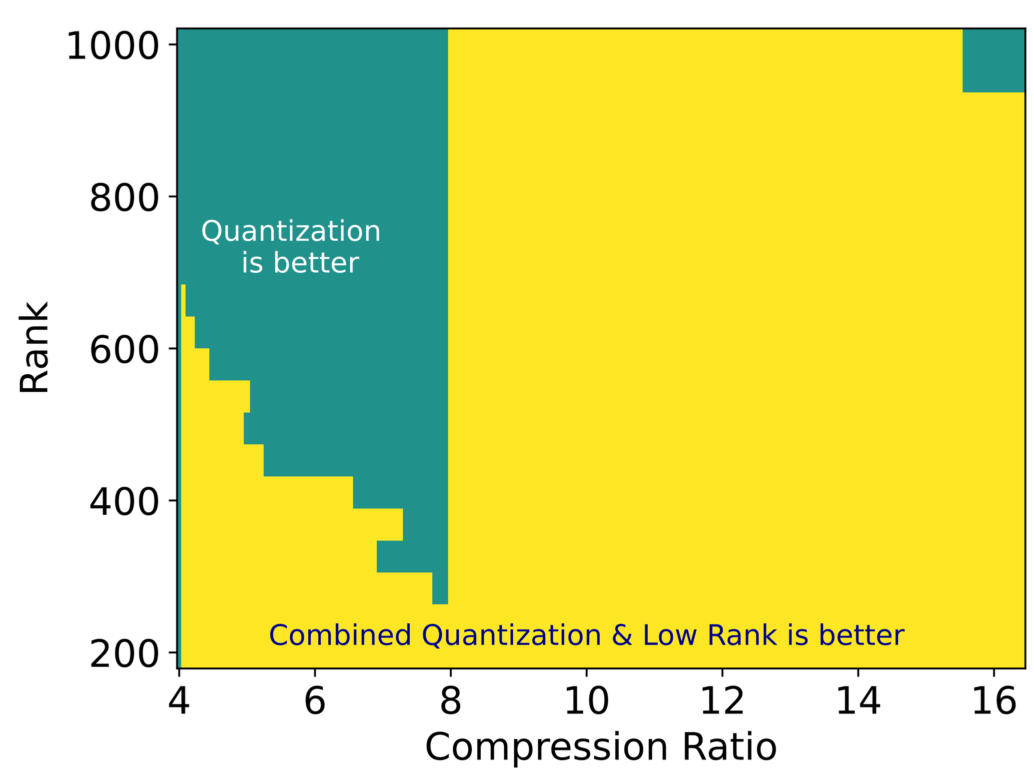
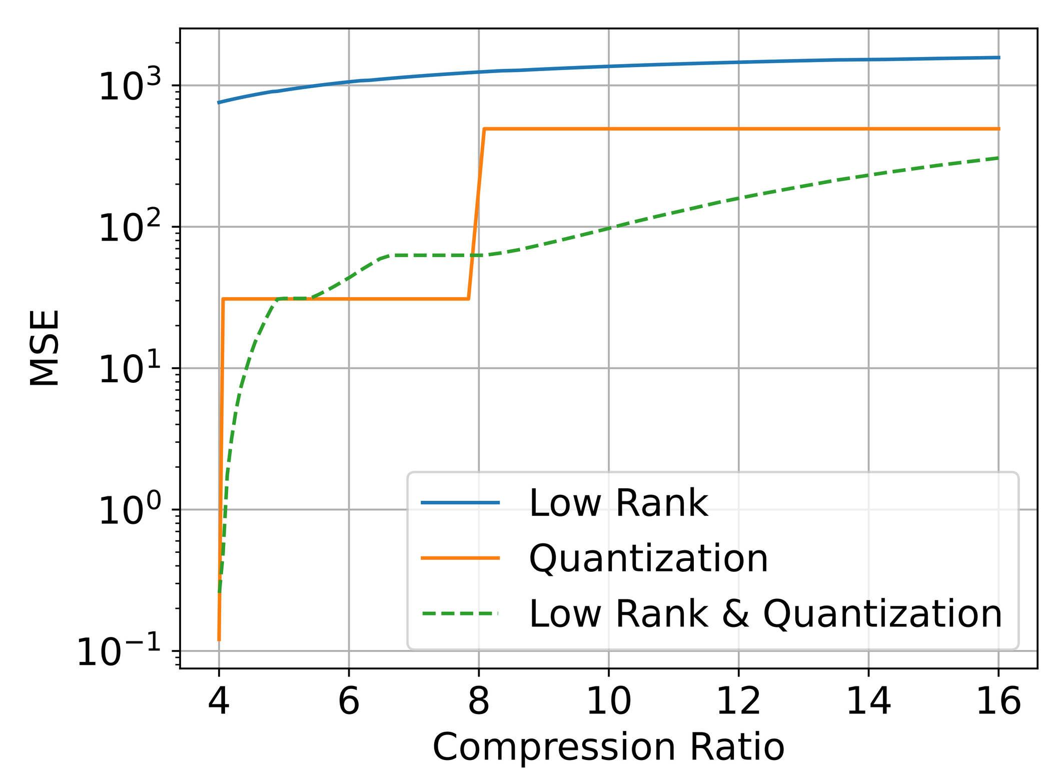


Table: Results on COCO dataset for object detection and instance segmentation.

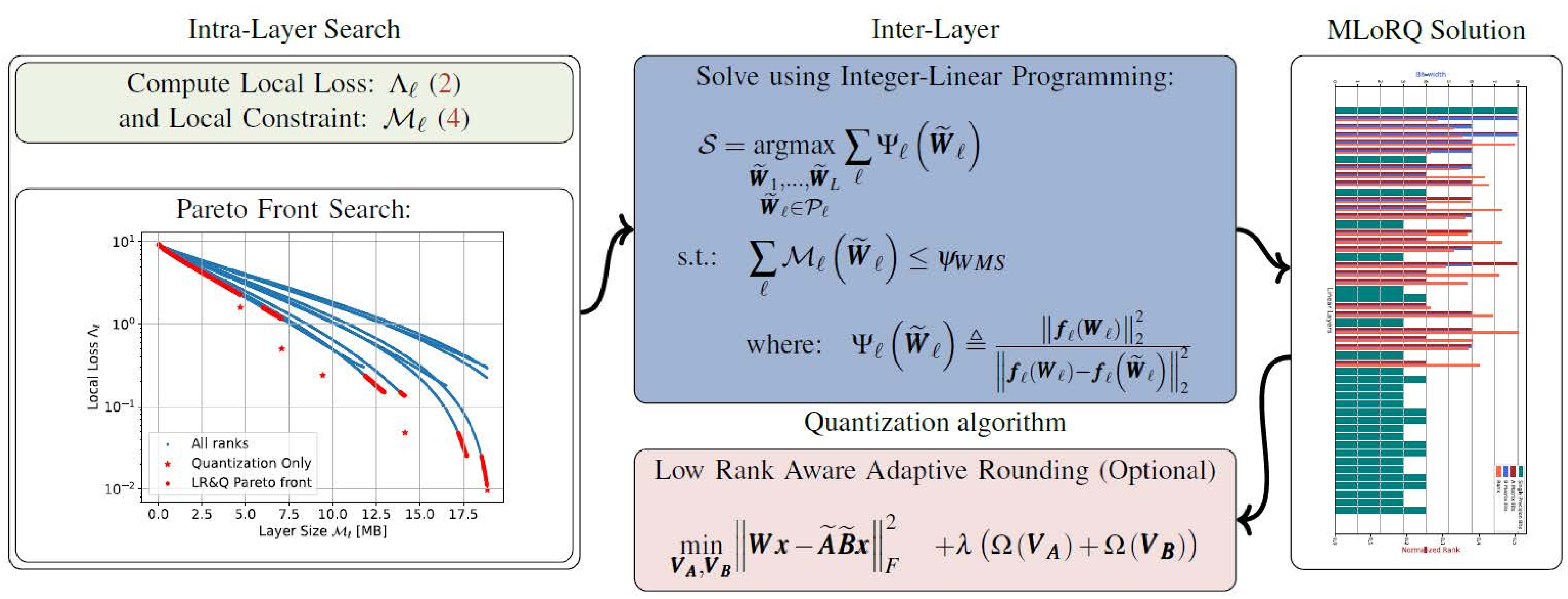
Method	Model Size (%)	Mask R-CNN w. Swin-T AP^{box}	Mask R-CNN w. Swin-T AP^{mask}	Mask R-CNN w. Swin-S AP^{box}	Mask R-CNN w. Swin-S AP^{mask}
Full-Precision	**100%**	**46.0**	**41.6**	**48.5**	**43.3**
GPTQ	9.3% (~3-bit)	22.9	25.0	31.7	32.5
RepQ-ViT		22.2	24.1	27.9	29.9
BRECQ		14.7	18.2	15.3	19.5
ERQ		27.2	28.9	30.6	33.0
MLoRQ		**33.7**	**32.9**	**36.8**	**35.5**

Why Joint Low-Rank & Quantization Optimization?





The MLoRQ Framework



Intra-Layer Search

Finds a set of candidate solutions for the ℓ^{th} layer, by minimizing the inner-layer optimization:

$$\min_{\boldsymbol{A}_\ell^{(r)}, \boldsymbol{B}_\ell^{(r)}, \phi_{\boldsymbol{A}_\ell}, \phi_{\boldsymbol{B}_\ell}} \Lambda_\ell\left(\widetilde{\boldsymbol{W}}_\ell\right) \quad \forall b_\ell^{\boldsymbol{A}}, b_\ell^{\boldsymbol{B}}, r_\ell$$

This procedure produces a **Pareto optimal** set of compression candidates, capturing the best trade-offs between memory footprint and approximation error, and significantly reducing the search space for later stages.

Inter-Layer Search

Finds a **mixed rank and quantization solution** $\mathcal{S}$ that assigns an optimal compression setup to each layer, minimizing the total error in the compressed model, using an Integer-Linear Programming formalization:

$$\mathcal{S} = \underset{\widetilde{\boldsymbol{W}}_1, \dots, \widetilde{\boldsymbol{W}}_L}{\text{argmax}} \sum_\ell \Psi_\ell\left(\widetilde{\boldsymbol{W}}_\ell\right)$$
$$\text{s.t.:} \quad \sum_\ell \mathcal{M}_\ell\left(\widetilde{\boldsymbol{W}}_\ell\right) \le \psi_{WMS},$$
$$\widetilde{\boldsymbol{W}}_\ell \in \mathcal{P}_\ell \quad \forall \ell$$

Results

Table: Top-1 accuracy (%) on ImageNet classification.

Model	Model Size (%)	ViT-S	ViT-B	DeiT-B	Swin-S
Full Prec.	100%	**81.39**	**84.54**	**81.8**	**83.23**
GPTQ	9.3% (~3-bit)	23.32	44.63	61.75	66.71
RepQ-ViT		15.65	26.98	58.92	59.83
ERQ		45.68	53.88	70.33	75.08
MLoRQ + RepQ		17.65	48.16	68.41	77.90
MLoRQ + ERQ		42.64	60.16	71.61	78.02
MLoRQ + ERQ + LRA		**46.55**	**69.56**	**76.12**	**79.27**
GPTQ	12.5% (~4-bit)	67.59	75.12	76.10	80.17
RepQ-ViT		65.05	68.48	75.61	79.45
ERQ		68.91	76.63	78.23	80.74
MLoRQ + RepQ		64.61	75.98	76.78	80.15
MLoRQ + ERQ		68.11	78.43	77.99	80.61
MLoRQ + ERQ + LRA		**69.12**	**79.13**	**78.61**	**80.76**

Come Visit Us!
https://github.com/ofirgo/MLoRQ