F-LBQ: Fine-grained Low Bit Quantization for Efficient and Accurate Object Detection —Suplementary Materials

February 6, 2025

1 Implementation Details

We randomly select 64 samples from training set to form the calibration set. We apply the proposed F-LBQ method on those finetuned ViT backbones to search for optimal bit-allocation for transformer layers that minimizes alteration on backbone feature caused by the quantization. For layers in detection and segmentation heads in Mask-RCNN framework, we apply standard uniform quantization and vanilla quantization scaling factors. We apply channel-group-wise allocation for F-LBQ experiments. For Lagrangian Multiplier Solver (F-LBQ-LG), we search the optimal λ that results in a bit-allocation b_l that is as close to the feasible region given by the bit-rate constraint as possible, with a tolerance ratio at $\left|\text{Bit}(q(\mathbf{\Phi}^{(1:L)})) - \mathcal{B}\right| \leq \epsilon$. In the experiments we set $\epsilon = 0.02$. Since we expect the slopes of the error curves for all layers w.r.t. different bit-widths are always positive, we operate the binary search on the power set $2^{\mathbb{R}}$, and set the searching range from $[2^{-100}, 0]$, which makes the asymptotic complexity of the binary search roughly $\Theta(\log \frac{100}{2\epsilon} L) \approx \Theta(96L)$. We observe the LG solver runs within one second on CPU for most of the time. To prevent large variance in layerwise bit-width from deteriorating the performance, we additionally constraint the minimum bit-width in the solution to $1 \sim 2$-bit lower than the target bit rate which is a hyperparameter.

2 More Discussions about Bit Allocation

The remarkable performance of our low-bit quanization method F-PTQ is thanks to the fine-grained multi-bit quantization as shown in Fig. 1. In the figure, we observe the two solvers behave differently mostly on earlier layers in the ViT detector, and both solvers allocate lowest bit rates to intermediate layers in the ViT backbone, implying that those layers contribute the least to the detection accuracy.

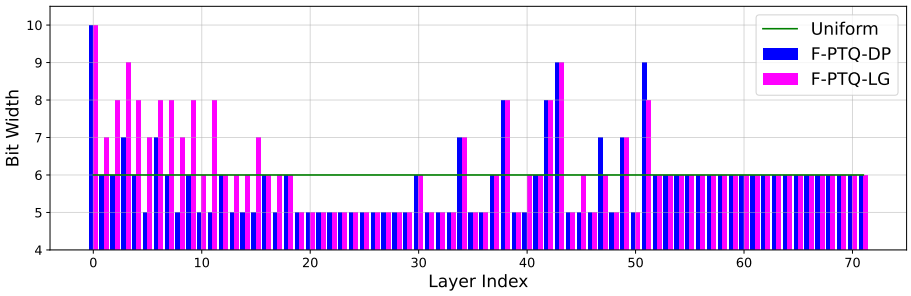

Figure 1: F-LBQ-DP v.s. F-LBQ-LG bit-allocation results on Mask-RCNN-Swin-T.