
Appendix for GPLQ: A General, Practical, and Lightning Quantization-Aware Training Method for Vision Transformers

Anonymous Author(s)

Affiliation

Address

email

A Appendix

This appendix provides supplementary experimental results and analyses that further support the findings presented in the main paper.

A.1 Detailed Performance with Varying PCA Dimensions

Due to space constraints in the main text, detailed results for downstream tasks across various datasets with different PCA dimensions were not fully elaborated. Table 1 presents these detailed outcomes for the Swin-T model under W4A4 quantization. Optimal results are highlighted in bold. As discussed in the main paper (Section 4.3, Table 6), the model generally achieves its best performance when the PCA dimension is set to 256, corresponding to a cumulative explained variance of approximately 61.3%.

Table 1: Detailed results of Swin-T(W4A4) on downstream tasks with different PCA dimensions used in the Act-QAT stage. ImageNet Top-1 accuracy (%) and cumulative explained variance (%) by the selected principal components are also shown. Best results on downstream tasks are highlighted in bold.

Setting	PCA dim	ImageNet (%)	Var. (%)	Downstream Task Accuracy (%)					Avg Task Acc (%)
				Aircraft	Food101	Flowers102	Pets	Cars	
w/o PCA	-	80.3	-	42.27	69.52	91.14	91.11	53.67	69.54
w/ PCA	64	80.4	24.1	42.57	70.37	92.36	91.99	54.74	70.41
w/ PCA	256	80.4	61.3	43.50	71.02	92.05	92.31	57.02	71.18
w/ PCA	512	80.3	87.7	42.51	70.25	92.00	92.67	56.01	70.69

Consistent with the conclusions in the main text, the Swin-T model demonstrates peak performance when the PCA dimension is 256, capturing around 60% of the cumulative variance. This configuration yields the best balance for generalization across the evaluated downstream tasks.

A.2 Quantization under Constrained Computational Resources

To assess the robustness and efficiency of our GPLQ method under more restrictive computational environments, we investigated the impact of varying the number of available GPUs during the activation quantization stage (Act-QAT). For these experiments, only activations were quantized for 1 epoch. The batch size per GPU was maintained at 16. Consequently, a reduction in the number of GPUs corresponds to a proportional decrease in the effective batch size and the learning rate was

20 adjusted accordingly. Other training parameters remained consistent with the settings described in
 21 the main paper.
 22 Table 2 details the performance of Swin-T on ImageNet and the associated training times.

Table 2: Impact of varying GPU counts on Swin-T (W4A4 Act-QAT only) ImageNet accuracy and training time. Batch size per GPU is 16.

Number of GPUs	Equivalent Batch Size	Learning Rate	ImageNet Acc (%)	Time (min)
8	128 (16×8)	5e-6	80.4	35
4	64 (16×4)	2.5e-6	80.3	71
2	32 (16×2)	1e-6	80.3	139
1	16 (16×1)	1e-6	80.3	275

23 The results indicate that GPLQ maintains high accuracy on ImageNet even as the number of GPUs
 24 and, consequently, the effective batch size and learning rate, are significantly reduced. The perfor-
 25 mance remains remarkably stable (80.3-80.4% top-1 accuracy) across all tested configurations. This
 26 resilience suggests that our method can effectively train quantized models even with minimal training
 27 resources, a capability not typically demonstrated by traditional QAT methods. Traditional QAT
 28 approaches often rely on large batch sizes and learning rates for stable convergence, and their perfor-
 29 mance is expected to degrade under such resource-constrained conditions. Our findings underscore
 30 the practical advantage of GPLQ in scenarios with limited hardware availability.

31 A.3 Impact of Training Data Volume on Model Performance

32 We further analyzed the influence of the training data volume on the Swin-T model’s performance
 33 during the 1-epoch activation quantization stage. All training settings were kept consistent with the
 34 GPLQ defaults, except for the number of training images used, which was varied by controlling the
 35 number of training iterations.

Table 3: Impact of training data volume (number of images used in 1 epoch of Act-QAT) on Swin-T ImageNet Top-1 accuracy (%) for W32A4 (only activations QAT) and W4A4 (activations QAT, weights PTQ) settings.

Training Iterations	Images Used	W32A4 Acc (%)	W4A4 Acc (%)
1	128	70.9	68.5
10	1,280	74.1	72.5
100	12,800	77.9	77.0
1000	128,000	79.9	79.2
10009	Full Dataset (approx. 1.28M)	80.4	79.8

36 As expected, increasing the volume of training data generally leads to improved model accuracy.
 37 Even with a relatively small number of images (e.g., 128,000, corresponding to about 10% of the full
 38 ImageNet training set for 1 epoch), the model achieves a respectable 79.2% accuracy. Training on
 39 the full dataset for 1 epoch yields 79.8% accuracy for the W32A4 model (output of Act-QAT stage),
 40 which forms a strong basis for the subsequent weight PTQ stage.

41 A.4 Impact of Direct W4A4 Training versus Sequential Quantization

42 To further highlight the benefits of our proposed sequential “activation-first, weights-later” (W32A4
 43 → W4A4) strategy, we compare it against a more direct approach where both weights and activations
 44 are quantized simultaneously from the start of the 1-epoch QAT process (direct W4A4). For this
 45 direct W4A4 baseline, all training aspects, including hyperparameters and the distillation method,
 46 were kept consistent with those used in the Act-QAT stage of our GPLQ method; the only difference
 47 lies in the quantization strategy (simultaneous QAT for W4A4 versus GPLQ’s sequential activation
 48 QAT followed by weight PTQ).

49 Table 4 presents the ImageNet top-1 accuracy for various Swin Transformer models.

Table 4: Comparison of ImageNet top-1 accuracy (%) for Swin Transformers using direct W4A4 QAT versus GPLQ’s sequential (W32A4 \rightarrow W4A4) approach. Optimal results are in **bold**.

Training Method	Swin-T	Swin-S	Swin-B	Swin-L
Direct W4A4	78.9	81.4	83.9	85.2
GPLQ (W32A4 \rightarrow W4A4)	79.8	81.9	84.2	85.5

50 The results clearly demonstrate that our proposed sequential quantization strategy (GPLQ) consistently
 51 outperforms the direct W4A4 QAT approach across all Swin Transformer variants. As discussed in
 52 the main paper, quantizing activations first (while keeping weights in FP32) and then applying PTQ to
 53 the weights offers several advantages. Beyond the slight improvements in training speed and reduced
 54 memory footprint during the Act-QAT stage (as pseudo-quantization of weights is not performed),
 55 this sequential approach helps to avoid the weight oscillations often encountered in traditional QAT.
 56 This leads to a smoother optimization process for the model weights, ultimately resulting in improved
 57 final model performance, as evidenced by the higher accuracies in Table 4.