LS-ViT: Least-Squares Hessian Based Block Reconstruction for Low-Bit Post-Training Quantization of Vision Transformers

Supplementary Material

A. Detailed Derivations

A.1. Derivation of Diagonal Least-Squares Hessian Approximation

For each index i, with N samples we have an overdetermined system. The LS objective is

$$\widehat{H}_{i,i} = \underset{H_{i,i}}{\operatorname{argmin}} \sum_{n=1}^{N} \left(g_i^{(n)} - H_{i,i}\Delta z_i^{(n)}\right)^2. \tag{23}$$

Differentiate w.r.t. $H_{i,i}$ and set to zero:

$$\frac{\partial}{\partial H_{i,i}} \sum_{n=1}^{N} \left(g_i^{(n)} - H_{i,i}\Delta z_i^{(n)}\right)^2 = 0, \tag{24}$$

$$-2\sum_{n=1}^{N} \left(g_i^{(n)} - H_{i,i}\Delta z_i^{(n)}\right) \Delta z_i^{(n)} = 0. \tag{25}$$

Rearranging gives

$$\sum_{n=1}^{N} g_i^{(n)}\Delta z_i^{(n)} = H_{i,i} \sum_{n=1}^{N} \left(\Delta z_i^{(n)}\right)^2, \tag{26}$$

hence

$$\widehat{H}_{i,i} = \frac{\sum_{n=1}^{N} g_i^{(n)}\Delta z_i^{(n)}}{\sum_{n=1}^{N}\left(\Delta z_i^{(n)}\right)^2} = \overline{g_i \Delta z_i} / \overline{(\Delta z_i)^2}. \tag{27}$$

The reconstruction loss follows as in Eq. (14).

A.2. Naive Rank-1 Least-Squares Solution

With $\mathbf{H}^{(\mathbf{z})} = \mathbf{u}\mathbf{u}^\top$,

$$\boldsymbol{g}^{(n)} = \mathbf{u}\mathbf{u}^\top \Delta\mathbf{z}^{(n)}. \tag{28}$$

The gradient of $\sum_{n=1}^{N} \|\boldsymbol{g}^{(n)} - \mathbf{u}\mathbf{u}^\top\Delta\mathbf{z}^{(n)}\|^2$ w.r.t. $\mathbf{u}$ is

$$\begin{aligned}
&\nabla_{\mathbf{u}} \sum_{n=1}^{N} \left\|\boldsymbol{g}^{(n)} - \mathbf{u}\mathbf{u}^\top\Delta\mathbf{z}^{(n)}\right\|^2 \\
&= -2\sum_{n=1}^{N} \left(\mathbf{u}^\top\Delta\mathbf{z}^{(n)}\mathbf{I} + \Delta\mathbf{z}^{(n)}\mathbf{u}^\top\right) \left(\boldsymbol{g}^{(n)} - \mathbf{u}\mathbf{u}^\top\Delta\mathbf{z}^{(n)}\right) \\
&= -2\sum_{n=1}^{N} \Big(\mathbf{u}^\top\Delta\mathbf{z}^{(n)}\,\boldsymbol{g}^{(n)} + \mathbf{u}^\top\boldsymbol{g}^{(n)}\,\Delta\mathbf{z}^{(n)} \\
&\qquad - \left(\mathbf{u}^\top\Delta\mathbf{z}^{(n)}\right)^2\mathbf{u} - \|\mathbf{u}\|^2\left(\mathbf{u}^\top\Delta\mathbf{z}^{(n)}\right)\Delta\mathbf{z}^{(n)}\Big)
\end{aligned} \tag{29}$$

Therefore, it is difficult to find a unique $\mathbf{u}$ that satisfies the Equation (29) due to nonlinear and high-order terms of $\mathbf{u}$.

A.3. Derivation of Low-Rank Least-Squares Hessian Approximation

Multiplying Equation (16) by $\Delta\mathbf{z}^{(n)\top}$ yields

$$\Delta\mathbf{z}^{(n)\top}\boldsymbol{g}^{(n)} = (\mathbf{u}^\top\Delta\mathbf{z}^{(n)})^2 \geq 0. \tag{30}$$

Since the pretrained full precision model is converged to a local minimum, we can assume the Hessian is positive-semidefinite (PSD) [13]:

$$\Delta\mathbf{z}^{(n)\top}\mathbf{H}^{(\mathbf{z})}\Delta\mathbf{z}^{(n)} \geq 0 \tag{31}$$

By Equation (6):

$$\Delta\mathbf{z}^{(n)\top}\mathbf{H}^{(\mathbf{z})}\Delta\mathbf{z}^{(n)} = \Delta\mathbf{z}^{(n)\top}\boldsymbol{g}^{(n)} = (\mathbf{u}^\top\Delta\mathbf{z}^{(n)})^2 \geq 0 \tag{32}$$

Since both vectors of $\mathbf{u}$ and $-\mathbf{u}$ define the same Hessian and thus the same reconstruction loss, we are free to choose the sign of $\mathbf{u}$:

$$\mathbf{H}^{(\mathbf{z})} = \mathbf{u}\mathbf{u}^\top = (-\mathbf{u})(-\mathbf{u})^\top \tag{33}$$

Without loss of generality, we can define $\mathbf{u}$ such that its projection onto $\Delta\mathbf{z}^{(n)}$ is non-negative, i.e.,

$$\mathbf{u}^\top\Delta\mathbf{z}^{(n)} \geq 0 \tag{34}$$

Equation (32) and (34) lead to the equation for each sample:

$$\mathbf{u}^\top\Delta\mathbf{z}^{(n)} = \sqrt{\Delta\mathbf{z}^{(n)\top}\boldsymbol{g}^{(n)}} \tag{35}$$

Therefore

$$\boldsymbol{g}^{(n)} = \mathbf{u}\sqrt{\Delta\mathbf{z}^{(n)\top}\boldsymbol{g}^{(n)}}. \tag{36}$$

The LS objective

$$\widehat{\mathbf{u}} = \min_{\mathbf{u}} \sum_{n=1}^{N} \left\|\boldsymbol{g}^{(n)} - \mathbf{u}\sqrt{\Delta\mathbf{z}^{(n)\top}\boldsymbol{g}^{(n)}}\right\|^2 \tag{37}$$

has gradient

$$-2\sum_{n=1}^{N} \sqrt{\Delta\mathbf{z}^{(n)\top}\boldsymbol{g}^{(n)}} \left(\boldsymbol{g}^{(n)} - \widehat{\mathbf{u}}\sqrt{\Delta\mathbf{z}^{(n)\top}\boldsymbol{g}^{(n)}}\right) = \mathbf{0}, \tag{38}$$

which rearranges to

$$\sum_{n=1}^{N} \boldsymbol{g}^{(n)}\sqrt{\Delta\mathbf{z}^{(n)\top}\boldsymbol{g}^{(n)}} = \widehat{\mathbf{u}}\sum_{n=1}^{N} \Delta\mathbf{z}^{(n)\top}\boldsymbol{g}^{(n)}. \tag{39}$$

Thus,

$$\widehat{\mathbf{u}} = \frac{\sum_{n=1}^{N} \boldsymbol{g}^{(n)}\sqrt{\Delta\mathbf{z}^{(n)\top}\boldsymbol{g}^{(n)}}}{\sum_{n=1}^{N} \Delta\mathbf{z}^{(n)\top}\boldsymbol{g}^{(n)}}, \tag{40}$$

and substituting into $\mathcal{L}_{\text{recon}} = \frac{1}{2}\Delta\mathbf{z}^\top\widehat{\mathbf{u}}\widehat{\mathbf{u}}^\top\Delta\mathbf{z}$ gives the expression reported in the main text.

Table 6. Ablation study of different Hessian approximation components on ImageNet. 'Approx.' denotes the approximation type: C (Constant, i.e., MSE), D (Diagonal), and L (Low-Rank). 'FM.' indicates the full-model computations for the approximation. We compare our diagonal (LS-ViT-D), low-rank (LS-ViT-L), and combined (LS-ViT) methods against various baselines. Best results are in **bold**.

Method	Approx.	FM.	W/A	ViT-S	ViT-B	DeiT-T	DeiT-S	DeiT-B	Swin-S	Swin-B
Full-Prec	-	-	32/32	81.39	84.54	72.71	79.85	81.80	83.23	85.27
MSE	C	-	3/3	41.05	74.75	46.88	50.95	72.97	74.67	76.57
BRECQ	D	1	3/3	54.33	66.62	49.27	63.72	72.82	75.20	77.64
APH	D	2	3/3	59.11	76.05	53.82	67.40	75.89	75.44	77.31
FIMA-Q-D	D	1	3/3	60.02	76.29	55.54	68.68	76.32	75.08	77.87
LS-ViT-D (ours)	D	1	3/3	63.25	77.35	55.55	69.30	76.55	77.29	79.03
FIMA-Q-L	L	15	3/3	64.09	77.46	55.25	68.91	76.33	76.03	77.59
LS-ViT-L (ours)	L	1	3/3	63.96	77.47	55.46	69.20	76.34	77.28	79.28
FIMA-Q	D+L	15	3/3	64.09	77.63	55.55	69.13	76.54	77.26	78.82
LS-ViT (ours)	D+L	1	3/3	**64.10**	**77.65**	**55.72**	**69.41**	**76.57**	**77.39**	**79.40**
QDrop	C	-	4/4	71.84	82.63	65.27	72.64	79.96	81.21	82.99
BRECQ	D	1	4/4	73.88	81.47	65.26	74.76	78.82	81.25	82.80
APH	D	2	4/4	75.63	82.10	66.25	76.23	80.14	81.40	83.11
FIMA-Q-D	D	1	4/4	75.88	83.02	66.81	76.79	80.19	81.18	83.35
LS-ViT-D (ours)	D	1	4/4	76.25	83.05	67.05	76.67	80.34	81.79	83.45
FIMA-Q-L	L	15	4/4	76.47	83.04	66.78	76.66	80.30	81.60	83.15
LS-ViT-L (ours)	L	1	4/4	76.62	83.08	66.94	76.67	80.40	81.80	83.62
FIMA-Q	D+L	15	4/4	76.65	83.04	66.84	76.87	80.33	**81.82**	83.60
LS-ViT (ours)	D+L	1	4/4	**76.67**	**83.08**	**67.05**	**76.89**	**80.41**	**81.82**	**83.62**

B. More Ablation Studies

B.1. Analysis of Each Component

To validate the effectiveness of our proposed components, we conduct a detailed ablation study on the ImageNet dataset, with results presented in Table 6. We analyze the performance of our diagonal-only (LS-ViT-D) and low-rank-only (LS-ViT-L) approximations and compare them against other reconstruction metrics, particularly the components of FIMA-Q. First, our LS-ViT-D consistently outperforms all other methods based on diagonal Hessian approximation, including BRECQ, APH, and FIMA-Q-D, at both 3/3 and 4/4 bits. This result confirms the superiority of our least-squares formulation in capturing the diagonal Hessian information more accurately than prior approaches. Second, we compare the low-rank components. Notably, our LS-ViT-L, which uses only a rank-1 approximation (as shown in the 'FM.' column), achieves better performance than FIMA-Q-L, which relies on a much higher-cost rank-15 approximation. This highlights the high efficiency and accuracy of our proposed low-rank estimation, which effectively approximates the representative Hessian without resorting to higher ranks. Finally, LS-ViT, which combines the diagonal and low-rank approximations, achieves the best overall performance. This demonstrates that our two proposed components are complementary and synergistically contribute to the state-of-the-art results.

Table 7. Ablation results (%) w.r.t. the sample size with W2/A3 and W3/A3 on ImageNet.

	Sample	W2/A3			W3/A3		
		ViT-S	DeiT-S	Swin-S	ViT-S	DeiT-S	Swin-S
FIMA-Q	**128**	14.59	38.85	40.01	49.64	64.12	71.71
	256	22.88	44.81	47.68	56.02	66.21	74.12
	512	30.91	48.11	54.12	60.45	67.87	75.80
	1024	38.82	52.57	59.57	64.09	69.13	77.26
LS-ViT	**128**	16.75	39.25	47.97	50.26	64.42	72.29
	256	26.18	45.28	56.26	56.83	66.45	74.62
	512	35.00	49.74	61.41	61.25	68.01	76.16
	1024	42.05	53.86	65.77	64.10	69.41	77.39

B.2. Effect of Calibration Set Size

We investigate the impact of the calibration sample size on quantization performance, with results shown in Table 7. We compare LS-ViT against FIMA-Q using sample sizes from 128 to 1024 under challenging low-bit (W2/A3 and W3/A3) settings. As expected, the performance of both methods improves as more calibration samples are used, which allows for a more stable and accurate estimation of the Hessian statistics. However, our LS-ViT consistently outperforms FIMA-Q across all tested sample sizes and models. The performance gap is particularly significant in the extremely low-bit W2/A3 scenario. For example, on Swin-S, LS-ViT with only 256 samples achieves 56.26% accuracy, surpassing FIMA-Q with 512 samples (54.12%). This demonstrates that our least-squares Hessian provides

Table 8. Top-1 accuracy of various ViT-based models on ImageNet for the W4/A4 setting. "*" denotes baselines differing only in the reconstruction metric; other settings are identical. "(-)" after APHQ-ViT shows results without MLP reconstruction for a metric-only comparison. **Bold** indicates the best top-1 accuracy.

Method	W/A	ViT-S	ViT-B	DeiT-T	DeiT-S	DeiT-B	Swin-S	Swin-B
Full-Prec	32/32	81.39	84.54	72.71	79.85	81.80	83.23	85.27
PTQ4ViT [40]	4/4	42.57	30.69	36.96	34.08	64.39	76.09	74.02
APQ-ViT [4]	4/4	47.95	41.41	47.94	43.55	67.48	77.15	76.48
RepQ-ViT [15]	4/4	65.05	68.48	57.43	69.03	75.61	79.45	78.32
ERQ [41]	4/4	68.91	76.63	60.29	72.56	78.23	80.74	82.44
IGQ-ViT [26]	4/4	73.61	79.32	62.45	74.66	79.23	80.98	83.14
GPTQ [7]	4/4	67.60	75.00	58.69	70.88	76.21	80.11	81.06
DuQuant [17]	4/4	1.73	9.53	22.68	26.57	64.87	78.59	78.56
SpinQuant [24]	4/4	50.41	5.42	42.57	58.44	60.67	64.16	50.80
OmniQuant [29]	4/4	8.80	5.50	10.67	16.85	38.74	77.48	78.14
SmoothQuant [37]	4/4	13.19	8.22	21.78	32.76	49.31	79.33	79.26
I&S-ViT [42]	4/4	74.87	80.07	65.21	75.81	79.97	81.17	82.60
DopQ-ViT [38]	4/4	75.69	80.95	65.54	75.84	80.13	81.71	83.34
OASQ [25]	4/4	72.88	76.59	66.31	76.00	78.83	81.02	82.46
QDrop* [32]	4/4	71.84	82.63	65.27	72.64	79.96	81.21	82.99
BRECQ* [13]	4/4	73.88	81.47	65.26	74.76	78.82	81.25	82.80
PD-Quant* [20]	4/4	74.48	82.54	66.95	76.62	80.18	81.64	83.47
APHQ-ViT [36]	4/4	76.07	82.41	66.66	76.40	80.21	81.81	83.42
APHQ-ViT(-)* [36]	4/4	75.63	82.10	66.25	76.23	80.14	81.40	83.11
FIMA-Q* [35]	4/4	76.65	83.04	66.84	76.87	80.33	**81.82**	83.60
LS-ViT (ours)	4/4	**76.67**	**83.08**	**67.05**	**76.89**	**80.41**	**81.82**	**83.62**

a more robust and data-efficient approximation, achieving superior results even with a limited number of calibration samples.

B.3. Extended W4/A4 Classification Results

As discussed in the main paper, we present a more comprehensive comparison of W4/A4 quantization results on ImageNet in Table 8. This table includes a wider array of recent SOTA methods, such as ERQ [41], DuQuant [17], SpinQuant [24], SmoothQuant [37], OmniQuant [29], and OASQ [25], providing a broader context for our performance. At the W4/A4 bit-width, the performance of SOTA methods is already highly optimized, making further improvements particularly challenging. As reported in FIMA-Q [35], increasing the computational budget (i.e., the rank) yields minimal gains at this bit-width. This suggests that the influence of the covariance term, which is critical in ultra-low-bit regimes, is comparatively less pronounced at 4-bit. Even in this context, LS-ViT still delivers consistent improvements. Compared to the strong FIMA-Q baseline, LS-ViT achieves an average accuracy improvement of 0.06%p across all seven models. This comprehensive comparison validates that LS-ViT is not only highly effective in ultra-low-bit scenarios (where the covariance term is crucial) but is also robustly effective in the W4/A4 setting, consistently outperforming the best prior work.

Table 9. Comparison with QAT under W3/A3 quantization.

Model	Method	Type	Data Size	Training Time	Acc.
DeiT-S	LSQ [6]	QAT	1280K	∼170 h	77.31
	QDrop [32]	PTQ	1024	105 min	50.95
	FIMA-Q [35]	PTQ	1024	225 min	69.13
	LS-ViT	PTQ	1024	105 min	69.41
Swin-S	LSQ [6]	QAT	1280K	∼450 h	80.62
	QDrop [32]	PTQ	1024	160 min	74.67
	FIMA-Q [35]	PTQ	1024	420 min	77.26
	LS-ViT	PTQ	1024	165 min	77.39

B.4. Comparison with QAT

To evaluate the practical upper bound, Table 9 compares LS-ViT with the QAT-based LSQ [6] under the W3/A3 setting. While QAT relies on extensive fine-tuning over hundreds of hours, LS-ViT achieves 77.39% accuracy on Swin-S in just 165 minutes, narrowing the performance gap to 3.23%p with only 1024 calibration images. These results demonstrate that LS-ViT provides highly comparable accuracy while being 160× faster than QAT, proving its effectiveness for resource-constrained scenarios.