FIMA-Q: Post-Training Quantization for Vision Transformers by Fisher Information Matrix Approximation

Supplementary Material

A. Main Proofs

A.1. Proof of Approximation 4

We believe that BRECQ uses the gradient of KL divergence instead of the task loss gradient is based on the following theorem:

Theorem A.1. *When the model's output distribution matches the true data distribution, the Hessian matrix of the KL divergence after a small perturbation of the model is exactly equal to the expectation of the Hessian matrix of the model's likelihood function.*

Proof. The Hessian matrix of the model's likelihood function is defined as:

$$\boldsymbol{H}(\theta) \triangleq \frac{\partial^2}{\partial\theta^2}\log f(X;\theta). \tag{22}$$

As mentioned in theorem 3.1, when the assumption that the model's output distribution matches the true data distribution is satisfied, the expectation of the Hessian matrix is equal to the negative Fisher Information Matrix.

We use the integral form of the KL divergence to derive the KL divergence after a small perturbation of the model. Assume the output distribution of the model is $p(x) = f(x;\theta)$, the output after perturbation is $q(x) = f(x;\theta')$, where θ' is a small perturbation w.r.t θ:

$$D_{\mathrm{KL}}(p\|q) = \int_{\mathbb{R}} f(x;\theta)\log\frac{f(x;\theta)}{f(x;\theta')}\,\mathrm{d}x. \tag{23}$$

Therefore, the Hessian matrix of KL divergence can be written as:

$$\frac{\partial^2}{\partial\theta'_i\partial\theta'_j}D_{\mathrm{KL}}(p\|q) = -\int_{\mathbb{R}} f(x;\theta)\left(\frac{\partial^2\log f(x;\theta')}{\partial\theta'_i\partial\theta'_j}\right)\mathrm{d}x. \tag{24}$$

It can be seen that when $f(x;\theta)$ matches the true data distribution, it can be regarded as the probability density function of the true data distribution. Thus, the Hessian matrix of KL divergence is equal to the expectation of the Hessian matrix of the log-likelihood of $f(x,\theta')$. When θ and θ' are sufficiently close, the Hessian matrix of the KL divergence is the expectation of the Hessian matrix of the model's log-likelihood function. □

A.2. Proof of Theorem 3.1

To prove Theorem 3.1, we begin with the definition of the score function and another theorem:

Definition 1. *The Fisher Information Matrix is defined as the variance of the score function, where the score function is the gradient of the log-likelihood function.*

Theorem A.2. *When the model's output distribution matches the true distribution, the expected value of the score function becomes* 0.

Proof. According to the definition of the score function, we have:

$$\begin{aligned}\mathbb{E}\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\middle|\theta\right] &= \int_{\mathbb{R}}\frac{\frac{\partial}{\partial\theta}f(x;\theta)}{f(x;\theta)}f(x;\theta)\mathrm{d}x\\ &= \frac{\partial}{\partial\theta}\int_{\mathbb{R}}f(x;\theta)\mathrm{d}x\\ &= \frac{\partial}{\partial\theta}1\\ &= 0,\end{aligned} \tag{25}$$

where the likelihood function $f(X;\theta)$ denotes the probability of the model output random variable X, and $f(x;\theta)$ denotes the probability density of X taking the value x. This equation holds if and only if the output distribution matches the true distribution, allowing the use of $f(x;\theta)$ as the probability density function for integration. □

Then, we can prove Theorem 3.1:

Proof. Based on the definition of the Fisher Information Matrix and the definition of variance $D(X) = E(X^2) - E^2(X)$, when the expected gradient of the log-likelihood is 0, we have:

$$\mathbf{F}(\theta) = \mathbb{D}\left[\frac{\partial}{\partial\theta}\log f(X;\theta)\middle|\theta\right] = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2\middle|\theta\right]. \tag{26}$$

The second derivative of the log-likelihood function with respect to the parameters (*i.e.*, the Hessian matrix) is:

$$\begin{aligned}&\frac{\partial^2}{\partial\theta^2}\log f(X;\theta) = \frac{\partial}{\partial\theta}\frac{\frac{\partial}{\partial\theta}f(X;\theta)}{f(X;\theta)}\\ &= \frac{\left(\frac{\partial^2}{\partial\theta^2}f(X;\theta)\right)\cdot f(X;\theta) - \left(\frac{\partial}{\partial\theta}f(X;\theta)\right)^2}{f(X;\theta)^2}\\ &= \frac{\frac{\partial^2}{\partial\theta^2}f(X;\theta)}{f(X;\theta)} - \left(\frac{\frac{\partial}{\partial\theta}f(X;\theta)}{f(X;\theta)}\right)^2\\ &= \frac{\frac{\partial^2}{\partial\theta^2}f(X;\theta)}{f(X;\theta)} - \left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2,\end{aligned} \tag{27}$$

where the second term is the definition of FIM, and the expectation of the first term is 0:

$$\mathbb{E}\left[\left.\frac{\frac{\partial^2}{\partial\theta^2}f(X;\theta)}{f(X;\theta)}\right|\theta\right]=\int_{\mathbb{R}}\frac{\frac{\partial^2}{\partial\theta^2}f(x;\theta)}{f(x;\theta)}f(x;\theta)\mathrm{d}x \\ =\frac{\partial^2}{\partial\theta^2}\int_{\mathbb{R}}f(x;\theta)\mathrm{d}x=0. \tag{28}$$

Therefore, when the model's output distribution matches the true distribution, the Fisher Information Matrix is equivalent to the expectation of the negative second derivative of the log-likelihood function, i.e., the expectation of the Hessian matrix of the negative log-likelihood function. □

A.3. Proof of Theorem 3.2

Proof. In the context of block-wise post-training quantization, we regard the KL divergence as a function of the perturbation to the block output:

$$\begin{aligned}\mathcal{L}_{\mathrm{KL}}(\Delta\mathbf{z}^{(b)})&=D_{\mathrm{KL}}(p(x;\mathbf{z}^{(b)})\|p(x;\mathbf{z}^{(b)}+\Delta\mathbf{z}^{(b)}))\\&=\int_{\mathbb{R}}p(x;\mathbf{z}^{(b)})\log\frac{p(x;\mathbf{z}^{(b)})}{p(x;\mathbf{z}^{(b)}+\Delta\mathbf{z}^{(b)})}\\&=\int_{\mathbb{R}}p(x;\mathbf{z}^{(b)})\log p(x;\mathbf{z}^{(b)})\\&\quad-\int_{\mathbb{R}}p(x;\mathbf{z}^{(b)})\log p(x;\mathbf{z}^{(b)}+\Delta\mathbf{z}^{(b)}).\end{aligned}\tag{29}$$

We perform a second order Taylor expansion to $\log p(x;\mathbf{z}^{(b)}+\Delta\mathbf{z}^{(b)})$ as below:

$$\begin{aligned}&\log p(x;\mathbf{z}^{(b)}+\Delta\mathbf{z}^{(b)})\\&=\log p(x;\mathbf{z}^{(b)})+\nabla_{\mathbf{z}^{(b)}}\log p(x;\mathbf{z}^{(b)})^\top\Delta\mathbf{z}^{(b)}\\&\quad+\frac{1}{2}\Delta\mathbf{z}^{(b)\top}\nabla^2_{\mathbf{z}^{(b)}}\log p(x;\mathbf{z}^{(b)})\Delta\mathbf{z}^{(b)}\end{aligned}\tag{30}$$

Thus, we have

$$\mathcal{L}_{\mathrm{KL}}(\Delta\mathbf{z}^{(b)})=-\Delta\mathbf{z}^{(b)\top}\cdot S_1-\frac{1}{2}\Delta\mathbf{z}^{(b)\top}\cdot S_2\cdot\Delta\mathbf{z}^{(b)},\tag{31}$$

where

$$\begin{aligned}S_1&=\int_{\mathbb{R}}p(x;\mathbf{z}^{(b)})\nabla_{\mathbf{z}^{(b)}}\log p(x;\mathbf{z}^{(b)})\\S_2&=\int_{\mathbb{R}}p(x;\mathbf{z}^{(b)})\nabla^2_{\mathbf{z}^{(b)}}\log p(x;\mathbf{z}^{(b)}).\end{aligned}\tag{32}$$

According to the properties of logarithmic function differentiation, we can deduce the following:

$$\nabla_{\mathbf{z}^{(b)}}p(x;\mathbf{z}^{(b)})=p(x;\mathbf{z}^{(b)})\nabla_{\mathbf{z}^{(b)}}\log p(x;\mathbf{z}^{(b)}).\tag{33}$$

Thus,

$$\begin{aligned}S_1&=\int_{\mathbb{R}}p(x;\mathbf{z}^{(b)})\nabla_{\mathbf{z}^{(b)}}\log p(x;\mathbf{z}^{(b)})\\&=\int_{\mathbb{R}}\nabla_{\mathbf{z}^{(b)}}p(x;\mathbf{z}^{(b)})\\&=\nabla_{\mathbf{z}^{(b)}}\int_{\mathbb{R}}p(x;\mathbf{z}^{(b)})\\&=\nabla_{\mathbf{z}^{(b)}}1\\&=0.\end{aligned}\tag{34}$$

For S_2, according to Eq. (33), the following equations hold:

$$\begin{aligned}&\nabla^2_{\mathbf{z}^{(b)}}p(x;\mathbf{z}^{(b)})\\&=\nabla_{\mathbf{z}^{(b)}}(p(x;\mathbf{z}^{(b)})\nabla_{\mathbf{z}^{(b)}}\log p(x;\mathbf{z}^{(b)}))\\&=\nabla_{\mathbf{z}^{(b)}}p(x;\mathbf{z}^{(b)})\nabla_{\mathbf{z}^{(b)}}\log p(x;\mathbf{z}^{(b)})^\top\\&\quad+p(x;\mathbf{z}^{(b)})\nabla^2_{\mathbf{z}^{(b)}}\log p(x;\mathbf{z}^{(b)})\\&=p(x;\mathbf{z}^{(b)})\nabla_{\mathbf{z}^{(b)}}\log p(x;\mathbf{z}^{(b)})\nabla_{\mathbf{z}^{(b)}}\log p(x;\mathbf{z}^{(b)})^\top\\&\quad+p(x;\mathbf{z}^{(b)})\nabla^2_{\mathbf{z}^{(b)}}\log p(x;\mathbf{z}^{(b)}).\end{aligned}\tag{35}$$

Therefore, S_2 can be written as

$$\begin{aligned}S_2&=\int_{\mathbb{R}}p(x;\mathbf{z}^{(b)})\nabla^2_{\mathbf{z}^{(b)}}\log p(x;\mathbf{z}^{(b)})\\&=\int_{\mathbb{R}}\nabla^2_{\mathbf{z}^{(b)}}p(x;\mathbf{z}^{(b)})\\&\quad-\int_{\mathbb{R}}p(x;\mathbf{z}^{(b)})\nabla_{\mathbf{z}^{(b)}}\log p(x;\mathbf{z}^{(b)})\nabla_{\mathbf{z}^{(b)}}\log p(x;\mathbf{z}^{(b)})^\top.\end{aligned}\tag{36}$$

According to the Leibniz'rule, we can derive that

$$\begin{aligned}\int_{\mathbb{R}}\nabla^2_{\mathbf{z}^{(b)}}p(x;\mathbf{z}^{(b)})&=\nabla^2_{\mathbf{z}^{(b)}}\int_{\mathbb{R}}p(x;\mathbf{z}^{(b)})\\&=\nabla^2_{\mathbf{z}^{(b)}}1\\&=0.\end{aligned}\tag{37}$$

Thus,

$$\begin{aligned}S_2&=-\int_{\mathbb{R}}p(x;\mathbf{z}^{(b)})\nabla_{\mathbf{z}^{(b)}}\log p(x;\mathbf{z}^{(b)})\nabla_{\mathbf{z}^{(b)}}\log p(x;\mathbf{z}^{(b)})^\top\\&=-\mathbf{F}^{(\mathbf{z}^{(b)})}.\end{aligned}\tag{38}$$

By substituting S_1 and S_2 into Eq. (31), we have:

$$\mathcal{L}_{\mathrm{KL}}(\Delta\mathbf{z}^{(b)})=\frac{1}{2}\Delta\mathbf{z}^{(b)\top}\mathbf{F}^{(\mathbf{z}^{(b)})}\Delta\mathbf{z}^{(b)}.\tag{39}$$

□

A.4. Derivation of Eq. (15)

Given:

$$\mathbf{F}^{(\mathbf{z}^{(b)})} = \boldsymbol{u}\boldsymbol{u}^\top, \tag{40}$$

$$\nabla\mathcal{L}_{\mathrm{KL}}(\Delta\mathbf{z}^{(b)}) = \mathbf{F}^{(\mathbf{z}^{(b)})}\Delta\mathbf{z}^{(b)}, \tag{41}$$

where $\boldsymbol{u}, \nabla\mathcal{L}_{\mathrm{KL}}(\Delta\mathbf{z}^{(b)}), \Delta\mathbf{z}^{(b)} \in \mathbb{R}^{a\times 1}$. We define a scalar $\alpha = \boldsymbol{u}^\top \cdot \Delta\mathbf{z}^{(b)}$ such that

$$\nabla\mathcal{L}_{\mathrm{KL}}(\Delta\mathbf{z}^{(b)}) = \alpha\boldsymbol{u}. \tag{42}$$

Then, we can deduce the below

$$\boldsymbol{u}^\top = \frac{\left(\nabla\mathcal{L}_{\mathrm{KL}}(\Delta\mathbf{z}^{(b)})\right)^\top}{\alpha}. \tag{43}$$

Thus,

$$\alpha = \sqrt{\nabla\mathcal{L}_{\mathrm{KL}}(\Delta\mathbf{z}^{(b),\top})\Delta\mathbf{z}^{(b)}}, \tag{44}$$

$$\boldsymbol{u} = \frac{\nabla\mathcal{L}_{\mathrm{KL}}(\Delta\mathbf{z}^{(b)})}{\sqrt{\nabla\mathcal{L}_{\mathrm{KL}}(\Delta\mathbf{z}^{(b),\top})\Delta\mathbf{z}^{(b)}}}. \tag{45}$$

A.5. Proof of Corollary 3.1

Proof. Given

$$\mathbf{F}^{(\mathbf{z}^{(b)})} = \boldsymbol{u}\boldsymbol{u}^\top, \tag{46}$$

$$\nabla\mathcal{L}_{\mathrm{KL}}(\Delta\mathbf{z}^{(b)}) = \mathbf{F}^{(\mathbf{z}^{(b)})}\Delta\mathbf{z}^{(b)}, \tag{47}$$

we can deduce the following

$$\Delta\mathbf{z}^{(b)\top}\nabla\mathcal{L}_{\mathrm{KL}}(\Delta\mathbf{z}^{(b)}) = \Delta\mathbf{z}^{(b)\top}\boldsymbol{u}\boldsymbol{u}^\top\Delta\mathbf{z}^{(b)}. \tag{48}$$

Since the right-hand side of Eq. (48) is a symmetric matrix, the left-hand side should also be symmetric:

$$\Delta\mathbf{z}^{(b)\top}\nabla\mathcal{L}_{\mathrm{KL}}(\Delta\mathbf{z}^{(b)}) = \left(\nabla\mathcal{L}_{\mathrm{KL}}(\Delta\mathbf{z}^{(b)})\right)^\top\Delta\mathbf{z}^{(b)}. \tag{49}$$

When $k = 1$, both sides of Eq. (49) are scalars, implying that Eq. (49) naturally holds. When $k > 1$, since $\Delta\mathbf{z}^{(b)}$ and $\mathcal{L}_{\mathrm{KL}}(\Delta\mathbf{z}^{(b)})$ are not directly related, we cannot guarantee their symmetry.

As a consequence, it is difficult to find a $\boldsymbol{u}$ such that $\mathbf{F}^{(\mathbf{z}^{(b)})} = \boldsymbol{u}\boldsymbol{u}^\top$ satisfying Eq. (10) in most cases. □

B. More Experiments

As both FIM approximation (FIMA) and reconstruction steps depend on calibration data, we separately evaluate their performance utilizing different number of samples. As shown in Table A, the accuracy using FIMA increases as sample size grows, but is generally robust to the sample size. However, the reconstruction step is more sensitive to the number of calibration samples.

Table A. Ablation results (%) w.r.t. the samples size with W3/A3 on ImageNet.

Sample Size	**In FIMA Step**			**In Reconstruction Step**		
	ViT-S	DeiT-S	Swin-S	ViT-S	DeiT-S	Swin-S
128	63.52	68.99	77.10	49.64	64.12	71.71
256	63.18	69.10	77.00	56.02	66.21	74.12
512	63.61	69.14	77.18	60.45	67.87	75.8
1024	64.09	69.13	77.26	64.09	69.13	77.26

Since the Fisher Information Matrix (FIM) captures global information, its computation involves averaging over the sample dimension. Theoretically, a larger sample size leads to a more accurate approximation due to reduced sampling error. However, since the averaging process mitigates the impact of individual sample variations, the difference is not particularly significant. In fact, even using a single sample for approximation can still yield an acceptable level of accuracy. However, as shown in Tab. A, directly altering the overall sample size leads to a more substantial accuracy change, as the reconstruction process in Adaround is more sensitive to the number of samples.