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

Zhuguanyu Wu[1,2*], Shihe Wang[1,2*], Jiayi Zhang[1,2], Jiaxin Chen[1,2], Yunhong Wang[1,2]

[1]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, China

[2]School of Computer Science and Engineering, Beihang University, Beijing, China

Introduction

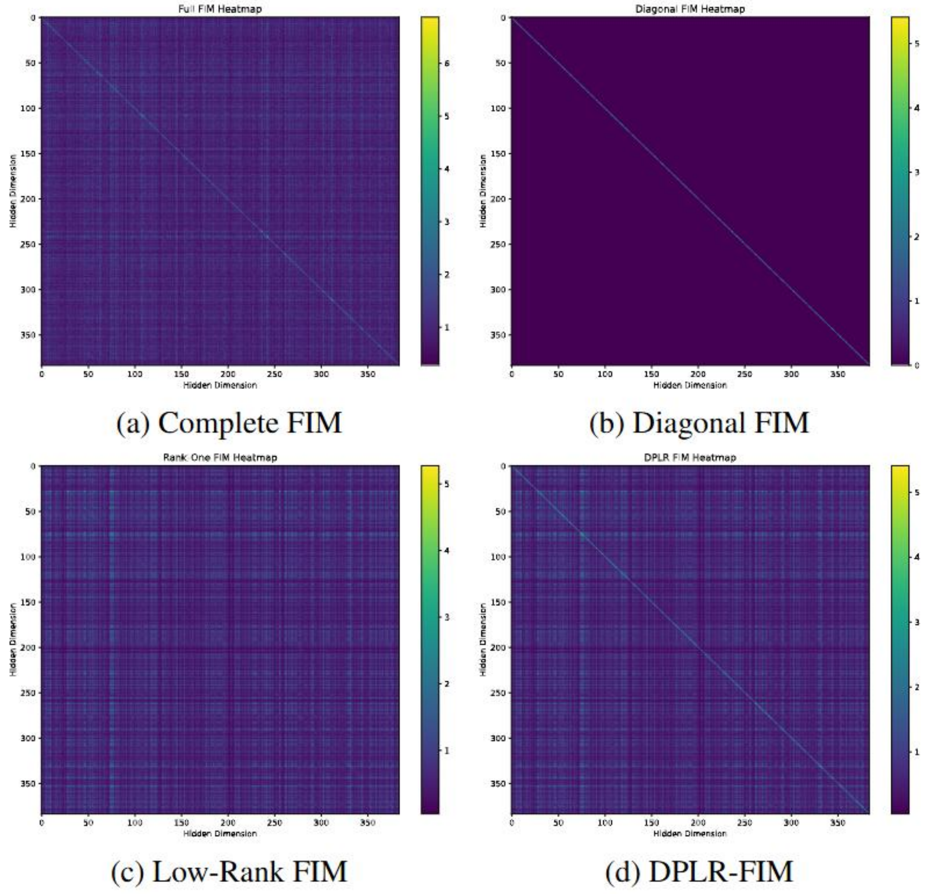
Problems

- **Diagonal approximation ignores the potentially useful off-diagonal information**
  - ➢ **Hessian-guided quantization loss:**

$$\min \left[-\Delta \mathbf{z}^{(b)} \ \mathbf{H}^{(\mathbf{z}^{(b)})} \Delta \mathbf{z}^{(b)} \right]$$

  - ➢ **All approximations in BRECQ:**

1. Using the Fisher information matrix (FIM) to approximate the negative Hessian matrix.
2. Using the **diagonal FIM** to approximate the full FIM.
3. Using the **square of the task loss gradient** to approximate the diagonal of the FIM.
4. Using the gradient of KL divergence instead of the task loss gradient.



(a) Complete FIM (b) Diagonal FIM
(c) Low-Rank FIM (d) DPLR-FIM

Problems

- **FIM is regarded as a sample-dependent matrix, and the second variance termis omitted**
  - **Hessian-guided quantization loss:**

$$\min \left[-\Delta \mathbf{z}^{(b)} \ \mathbf{H}^{(\mathbf{z}^{(b)})} \Delta \mathbf{z}^{(b)} \right]$$

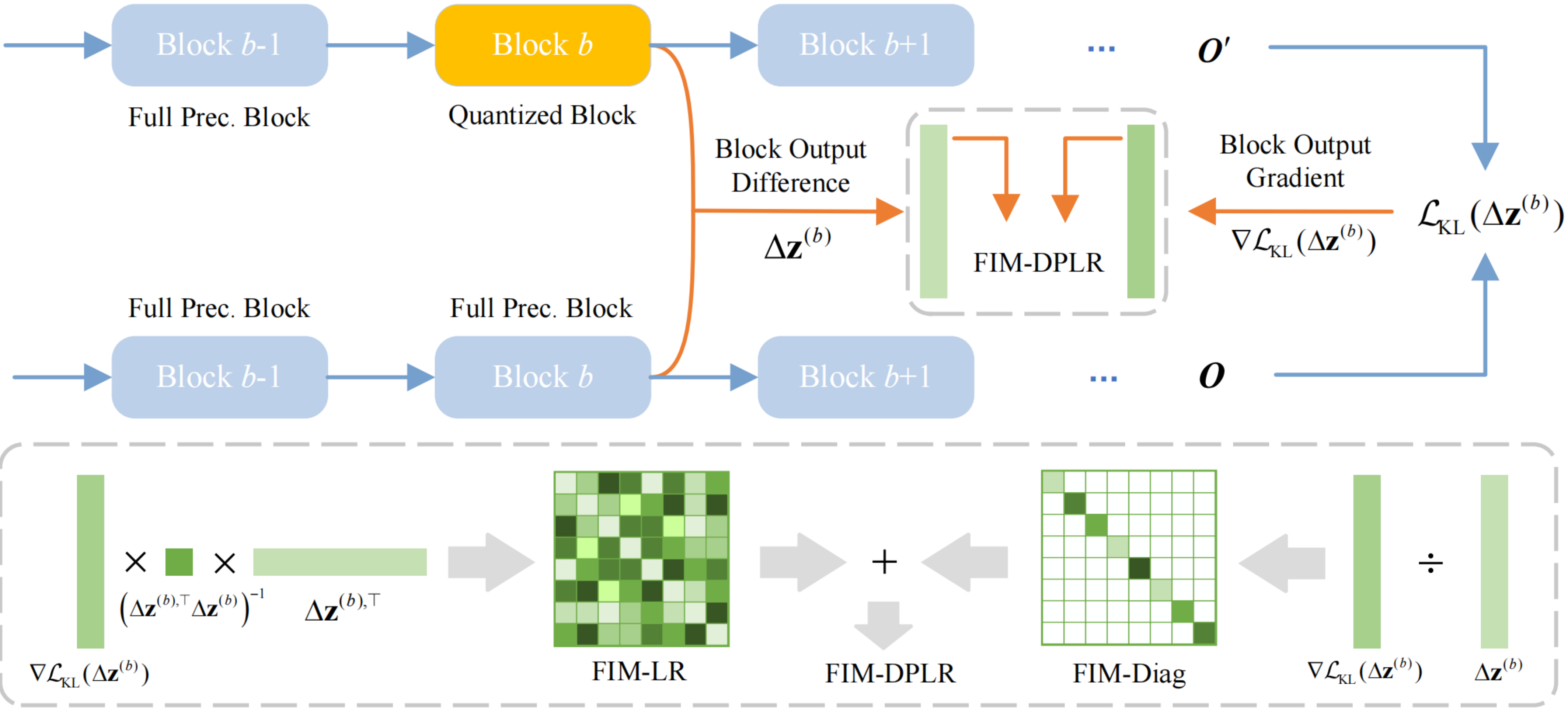
  - **All approximations in BRECQ:**

1. Using the Fisher information matrix (FIM) to approximate the negative Hessian matrix.
2. Using the **diagonal FIM** to approximate the full FIM.
3. Using the **square of the task loss gradient** to approximate the diagonal of the FIM.
4. Using the gradient of KL divergence instead of the task loss gradient.

$$\begin{aligned}\mathrm{Diag}\left(\mathbf{F}^{(\mathbf{z}^{(b)})}\right) &= \left[\left(\nabla_{\mathbf{z}^{(b)}} \log p(y;\mathbf{z}^{(b)})\right)^2\right] \\ &= \left[\left(\nabla_{\mathbf{z}^{(b)}} \log p(y;\mathbf{z}^{(b)})\right)\right]^2 \\ &\quad \boxed{+\mathrm{Var}\left(\nabla_{\mathbf{z}^{(b)}} \log p(y;\mathbf{z}^{(b)})\right)}\end{aligned}$$

Method

Overview



Relationship Between FIM and KL

- **FIM is linearly proportional to the gradient of KL divergence.**

$$\mathbf{F}^{(\mathbf{z}^{(b)})} = \mathrm{E}\left[\left(\nabla_{\mathbf{z}^{(b)}} \log p(y;\mathbf{z}^{(b)})\right)\left(\nabla_{\mathbf{z}^{(b)}} \log p(y;\mathbf{z}^{(b)})\right)^{\cdot}\right],$$

$$\mathrm{L}_{\mathrm{KL}}(\Delta\mathbf{z}^{(b)}) = D_{\mathrm{KL}}(p(y;\mathbf{z}^{(b)}) \| p(y;\mathbf{z}^{(b)}+\Delta\mathbf{z}^{(b)})) = \sum_{x} p(y;\mathbf{z}^{(b)}) \log \frac{p(y;\mathbf{z}^{(b)})}{p(y;\mathbf{z}^{(b)}+\Delta\mathbf{z}^{(b)})}.$$

$${}_{\mathrm{KL}}(\Delta\mathbf{z}^{(b)}) = \frac{1}{2}\Delta\mathbf{z}^{(b),}\ \mathbf{F}^{(\mathbf{z}^{(b)})}\Delta\mathbf{z}^{(b)}$$

$$\nabla\ {}_{\mathrm{KL}}(\Delta\mathbf{z}^{(b)}) = \frac{\partial\ {}_{\mathrm{KL}}(\Delta\mathbf{z}^{(b)})}{\partial\Delta\mathbf{z}^{(b)}} = \mathbf{F}^{(\mathbf{z}^{(b)})}\Delta\mathbf{z}^{(b)}$$

IRIP laboratory

Approximations of FIM

- **The rank-one approximation(FIM-OR)：**
  - We set： $\mathbf{F}^{(\mathbf{z}^{(b)})}_{\mathrm{rank}-1} = uu^{\bullet}\ ,\ \ u \in \mathbb{R}^{a\times 1}$
  - Based on the relationship between the Fisher information matrix and the KL divergence gradient, we have：

$$u = \frac{\nabla \quad_{\mathrm{KL}}(\Delta\mathbf{z}^{(b)})}{\sqrt{\nabla \quad_{\mathrm{KL}}(\Delta\mathbf{z}^{(b)}) \quad \Delta\mathbf{z}^{(b,i)}}}$$

  - Hessian-guided quantization loss：

$$_{\mathrm{rank}-1} = \frac{\left(\Delta\mathbf{z}^{(b,i)} \quad \nabla \quad_{\mathrm{KL}}(\Delta\mathbf{z}^{(b)})\right)^2}{\nabla \quad_{\mathrm{KL}}(\Delta\mathbf{z}^{(b)}) \quad \Delta\mathbf{z}^{(b,i)}}$$

- **The low-rank approximation (FIM-LR)：**
  - The Moore-Penrose inverse of $\Delta\mathbf{z}^{(b)}$:

$$\Delta\mathbf{z}^{(b),+} = \left(\Delta\mathbf{z}^{(b)\text{晻}}\Delta\mathbf{z}^{(b)}\right)^{-1} \Delta\mathbf{z}^{(b)}$$

  - Hessian-guided quantization loss：

$$_{\mathrm{rank}-k} = \Delta\mathbf{z}^{(bi)} \quad \mathbf{F}^{(\mathbf{z}^{(b)})}\Delta\mathbf{z}^{(b,i)} = A\cdot B\cdot C$$

$$A = \Delta\mathbf{z}^{(b,i)} \quad \nabla \quad_{\mathrm{KL}}(\Delta\mathbf{z}^{(b)}),$$

$$B = \left(\Delta\mathbf{z}^{(b)} \quad \Delta\mathbf{z}^{(b)}\right)^{-1},$$

$$C = \Delta\mathbf{z}^{(b)} \quad \Delta\mathbf{z}^{(b,i)}.$$

Method

Approximations of FIM

- **The diagonal approximation (FIM-Diag):**

$$\mathbf{F}_{\mathrm{Diag}}^{(\mathbf{z}^{(b)})} = \mathrm{Diag}\left(\frac{\nabla \mathrm{L}_{\mathrm{KL}}(\Delta\mathbf{z}^{(b)})_1}{\Delta\mathbf{z}_1^{(b)}}, \cdots, \frac{\nabla \mathrm{L}_{\mathrm{KL}}(\Delta\mathbf{z}^{(b)})_a}{\Delta\mathbf{z}_a^{(b)}}\right),$$

$$\mathrm{L}_{\mathrm{diag}} = \left(\frac{\nabla \mathrm{L}_{\mathrm{KL}}(\Delta\mathbf{z}^{(b)})}{\Delta\mathbf{z}^{(b)}}\right)^{\cdot} \left(\Delta\mathbf{z}^{(b,i)}\right)^2$$

- **The diagonal plus low-rank approximation (FIM-DPLR):**

$$\mathbf{F}_{\mathrm{DPLR}}^{(\mathbf{z}^{(b)})} = \alpha \cdot \mathbf{F}_{\mathrm{rank}-k}^{(\mathbf{z}^{(b)})} + (1-\alpha) \cdot \mathbf{F}_{\mathrm{diag}}^{(\mathbf{z}^{(b)})}$$

$$\mathrm{L}_{\mathrm{DPLR}} = \alpha \cdot \mathrm{L}_{\mathrm{rank}-k} + (1-\alpha) \cdot \mathrm{L}_{\mathrm{diag}}$$

Algorithm 1 Pipeline of Block-wise FIMA-Q.

Input: Full-precision model $\mathcal{M}$, full-precision block $\mathcal{B}_{\mathrm{full}}$, quantized block $\mathcal{B}_{\mathrm{quant}}$, calibration data $\mathcal{D}_{\mathrm{calib}}$, rank r, maximal iteration number max_iter, and iteration interval x.

Output: The optimized quantized block $\mathcal{B}_{\mathrm{quant}}$.

\# Calculate Rank-One FIM:

Generate the raw input $\boldsymbol{X}_{\mathrm{raw}}$ and output $\mathbf{z}^{(b)}$ of $\mathcal{B}_{\mathrm{full}}$, the quantized input $\boldsymbol{X}_{\mathrm{quant}}$ of $\mathcal{B}_{\mathrm{quant}}$ based on $\mathcal{D}_{\mathrm{calib}}$.
Generate the model output O_{raw} and O_{quant} by performing forward propagations through the network starting from $\mathcal{B}_{\mathrm{raw}}$ and $\mathcal{B}_{\mathrm{quant}}$ based on $\boldsymbol{X}_{\mathrm{raw}}$, respectively.
Calculate the perturbation $\Delta\mathbf{z}^{(b)}$ and the loss $\mathcal{L}_{\mathrm{KL}}(\Delta\mathbf{z}^{(b)})$ based on Eq. (9), and perform backward propagation to compute the gradient $\nabla\mathcal{L}_{\mathrm{KL}}(\Delta\mathbf{z}^{(b)})$.

\# Perform Quantization Reconstruction:

Initialize the current rank $k = 1$, and the next data iteration $y = x$.
for $i = 1, \cdots,$ max_iter **do**
    if $k < r$ **and** $i = y$ **then**
        Calculate $\Delta\mathbf{z}^{(b)\prime}$ and $\nabla\mathcal{L}_{\mathrm{KL}}(\Delta\mathbf{z}^{(b)\prime})$.
        Calculate B in Eq. (19).
        Set $k := k + 1$ and $y := y + x$.
    end if
    Calculate $\mathcal{L}_{\mathrm{DPLR}}$ with Eq. (21) and perform BP.
    Update all the AdaRound weights by [24] and activation scaling factors in $\mathcal{B}_{\mathrm{quant}}$.
end for

Experiment

Main Results

- **Results on ImageNet**

Method	OP	SQ	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	×	✓	3/3	0.10	0.10	3.50	0.10	31.06	28.69	20.13
RepQ-ViT	×	✓	3/3	0.10	0.10	0.10	0.10	0.10	0.10	0.10
AdaLog	×	✓	3/3	13.88	37.91	31.56	24.47	57.47	64.41	69.75
I&S-ViT	✓	✓	3/3	45.16	63.77	41.52	55.78	73.30	74.20	69.30
DopQ-ViT	✓	✓	3/3	54.72	65.76	44.71	59.26	74.91	74.77	69.63
QDrop*	✓	×	3/3	41.05	74.75	46.88	50.95	72.97	74.67	76.57
FIMA-Q (Ours)	✓	×	3/3	**64.09**	**77.63**	**55.55**	**69.13**	**76.54**	**77.26**	**78.82**
PTQ4ViT	×	✓	4/4	42.57	30.69	36.96	34.08	64.39	76.09	74.02
APQ-ViT	×	✓	4/4	47.95	41.41	47.94	43.55	67.48	77.15	76.48
RepQ-ViT	×	✓	4/4	65.05	68.48	57.43	69.03	75.61	79.45	78.32
ERQ	×	✓	4/4	68.91	76.63	60.29	72.56	78.23	80.74	82.44
IGQ-ViT	×	✓	4/4	73.61	79.32	62.45	74.66	79.23	80.98	83.14
AdaLog	×	✓	4/4	72.75	79.68	63.52	72.06	78.03	80.77	82.47
I&S-ViT	✓	✓	4/4	74.87	80.07	65.21	75.81	79.97	81.17	82.60
DopQ-ViT	✓	✓	4/4	75.69	80.95	65.54	75.84	80.13	81.71	83.34
QDrop*	✓	×	4/4	71.84	82.63	65.27	72.64	79.96	81.21	82.99
OASQ	✓	×	4/4	72.88	76.59	66.31	76.00	78.83	81.02	82.46
FIMA-Q (Ours)	✓	×	4/4	**76.68**	**83.04**	**66.84**	**76.87**	**80.33**	**81.82**	**83.60**

Experiment

Main Results

- **Results on COCO**

Method	Opt	SQ	W/A	Mask R-CNN Swin-T AP^b	Mask R-CNN Swin-T AP^m	Mask R-CNN Swin-S AP^b	Mask R-CNN Swin-S AP^m	Cascade Mask R-CNN Swin-T AP^b	Cascade Mask R-CNN Swin-T AP^m	Cascade Mask R-CNN Swin-S AP^b	Cascade Mask R-CNN Swin-S AP^m
Full-Precision	-	-	32/32	46.0	41.6	48.5	43.3	50.4	43.7	51.9	45.0
Baseline	×	×	4/4	34.6	34.2	40.8	38.6	45.9	40.2	47.9	41.6
PTQ4ViT	×	✓	4/4	6.9	7.0	26.7	26.6	14.7	13.5	0.5	0.5
APQ-ViT	×	✓	4/4	23.7	22.6	**44.7**	40.1	27.2	24.4	47.7	41.1
RepQ-ViT	×	✓	4/4	36.1	36.0	$44.2_{42.7}$†	40.2	47.0	41.1	49.3	43.1
ERQ	×	✓	4/4	36.8	36.6	43.4	40.7	47.9	42.1	50.0	43.6
I&S-ViT	✓	✓	4/4	37.5	36.6	43.4	40.3	48.2	42.0	50.3	43.6
DopQ-ViT	✓	✓	4/4	37.5	36.5	43.5	40.4	48.2	42.1	50.3	**43.7**
QDrop*	✓	×	4/4	36.2	35.4	41.6	39.2	47.0	41.3	49.0	42.5
FIMA-Q (Ours)	✓	×	4/4	**38.7**	**37.8**	44.2	**41.1**	**48.7**	**42.5**	**50.4**	**43.7**

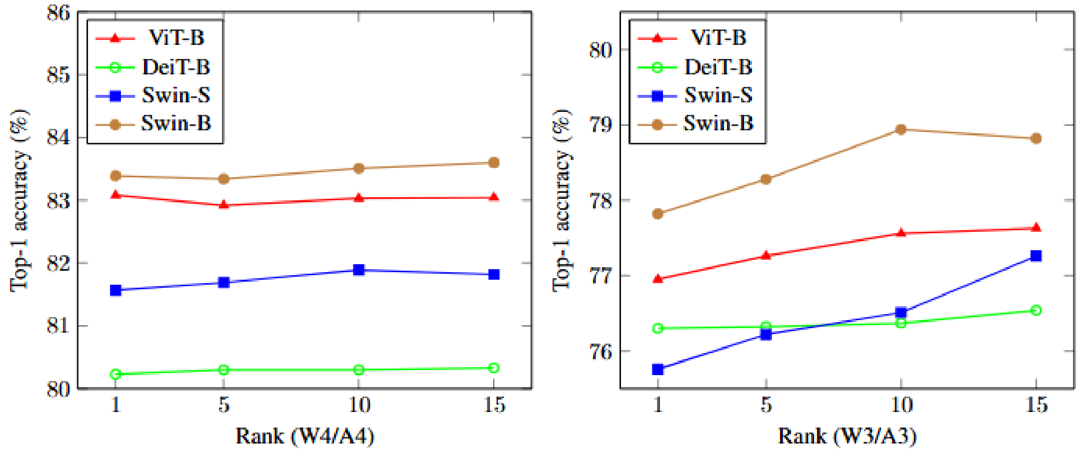
Experiment

Ablation Study

- **Comparison of different quantization losses**

#Bits (W/A)	Method	ViT-S	ViT-B	DeiT-T	DeiT-S	DeiT-B	Swin-S	Swin-B
4/4	MSE	71.84	82.63	65.27	72.64	79.96	81.21	82.99
	BRECQ-FIM	63.70	76.26	61.99	72.52	76.59	80.52	81.80
	Diag-FIM (**Ours**)	75.88	83.02	66.81	76.79	80.19	81.18	83.35
	LR-FIM (**Ours**)	76.47	**83.04**	66.78	76.66	80.30	81.60	83.15
	DPLR-FIM (Ours)	**76.65**	**83.04**	**66.84**	**76.87**	**80.33**	**81.82**	**83.60**

- **Comparison of different ranks**



- **The training time cost (GPU Minutes) using different ranks**

Model	Qdrop*	Ours			
		k=1	k=5	k=10	k=15
DeiT-T	100	100	115	150	180
DeiT-S	105	105	120	160	225
DeiT-B	145	150	160	240	310
Swin-S	160	165	235	360	420
Swin-B	170	180	250	420	480

Conclusion

- In this paper, we propose a novel approach dubbed FIMA-Q for post-training quantization of Vision Transformers. Specifically, we propose a more accurate approximation method for FIM.
- Specifically, we first demonstrate that FIM is proportional to the gradient of KL-divergence, based on which we develop a novel estimation method dubbed DPLR-FIM by integrating both the diagonal and off-diagonal information.
- Extensive experimental results across distinct vision transformer architectures validate the effectiveness of FIMA-Q for various visual tasks. The results reveal that our method has achieved significant performance improvements in the cases of low-bit quantization, especially with an average improvement of 5.31%, compared to the current state-of-the-art approaches in 3-bit quantization.

北京航空航天大学计算机学院

School of Computer Science and Engineering, Beihang University

Thanks for your listening