CVPR Nashville JUNE 11-15, 2025

IRIP laboratory

APHQ-ViT: Post-Training Quantization with Average Perturbation Based Reconstruction for Vision Transformers

Zhuguanyu Wu[1,2], Jiayi Zhang[1,2], Jiaxin Chen[1,2]✉, Jinyang Guo[3], Di Huang[2], Yunhong Wang[1,2] ✉

虚拟现实技术与系统全国重点实验室
STATE KEY LABORATORY OF VIRTUAL REALITY TECHNOLOGY AND SYSTEMS

北京航空航天大学计算机学院
School of Computer Science and Engineering, Beihang University

人工智能学院(人工智能研究院)
School of Artificial Intelligence (Institute of Artificial Intelligence)

Introduction

Background

- Model quantization converts the weights and activations from floating-precision to low bit-width integers. Reconstruction-based Post-Training Quantization (PTQ) method rely solely on a **small unlabeled** dataset, and achieves superior accuracy by introducing an efficient fine-tuning process.

- Recent reconstruction-based PTQ methods suffer from the two limitations. **1) Inaccurate estimation of output importance**. **2) Performance degradation in quantizing post-GELU activation**.

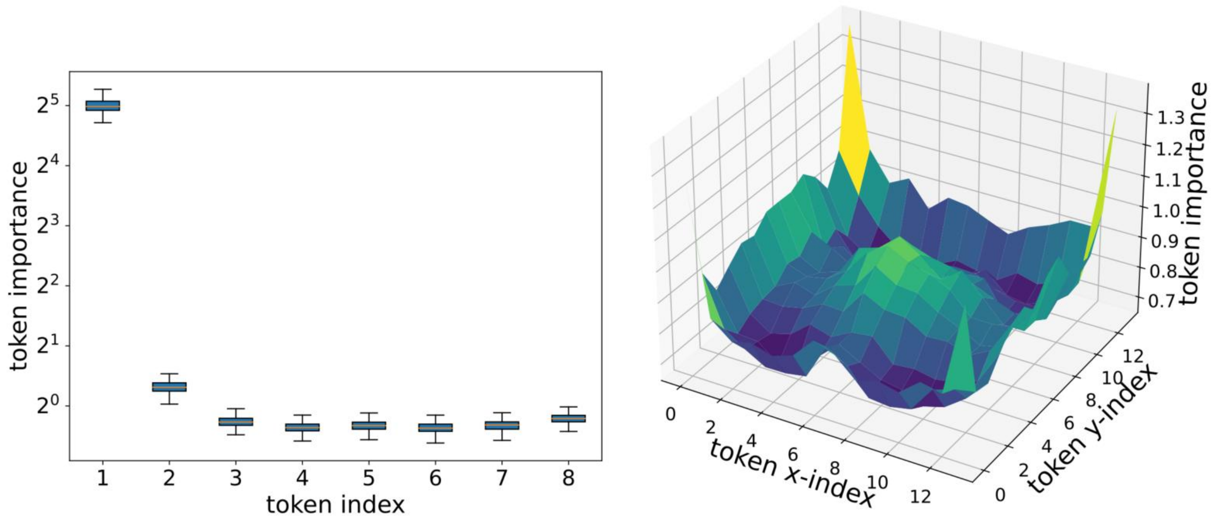
  - 1) Inaccurate estimation of output importance. Existing methods use MSE or FIM based quantization loss during block reconstruction, which is suboptimal.

  - 2) Performance degradation in quantizing post-GELU activation. The activation range reaching up to 40 in certain layers leads to an unstable fine-tuning process.

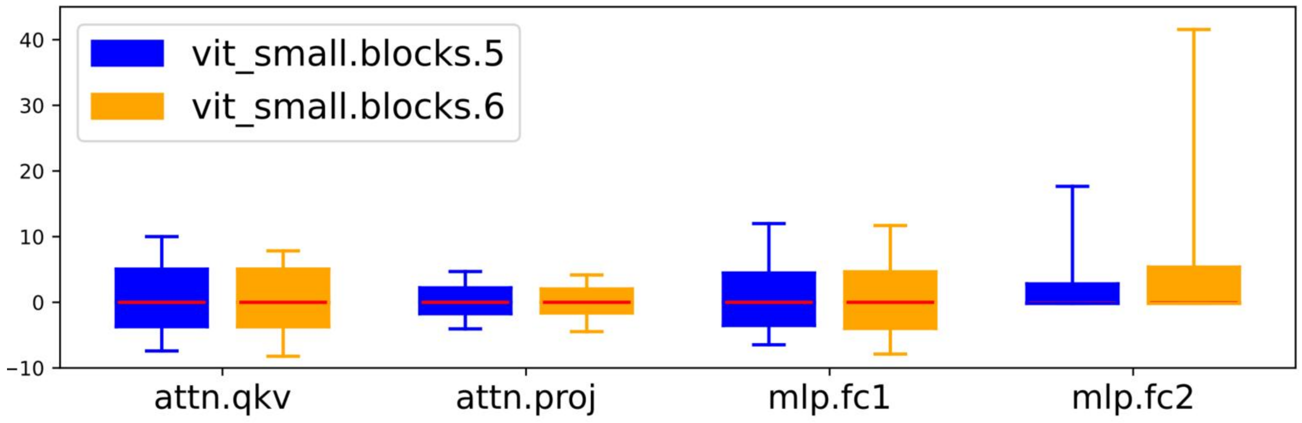
Introduction

Observations

- The importance of the class token (*i.e.*, the first token) is much higher than that of the patch tokens, and distinct patch tokens also have substantially different APH importance.



Token Importance

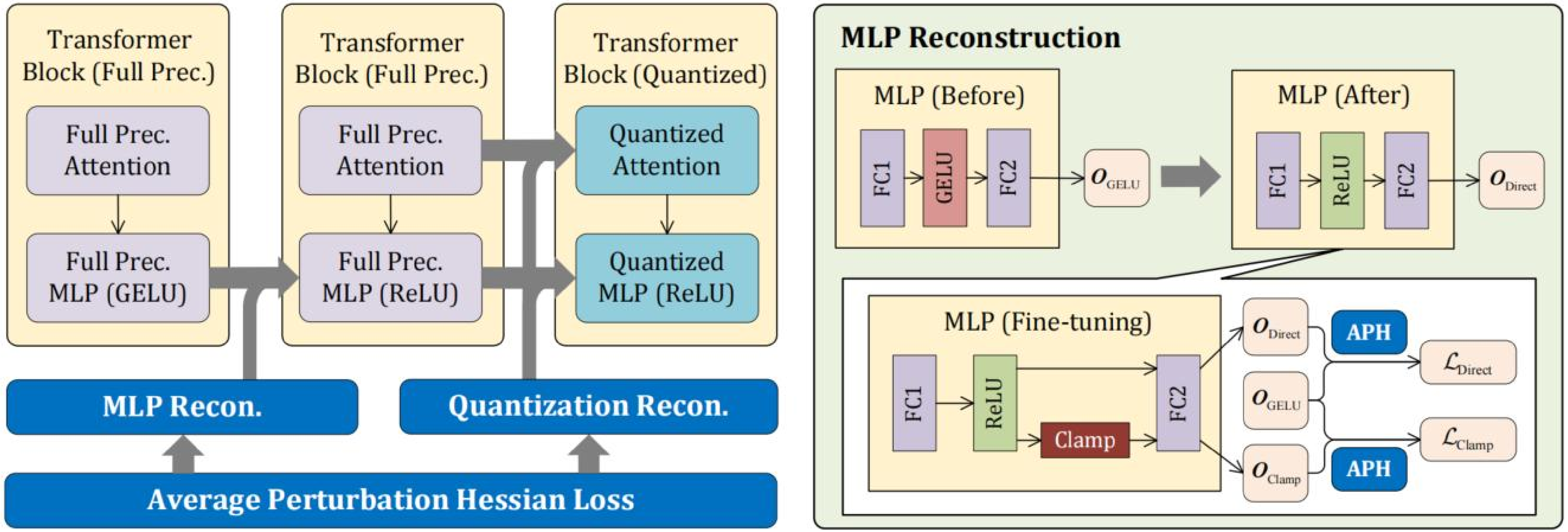


Activation statistics

- The activation range reaching up to 40 in certain post-GELU layers (*i.e.*, mlp.fc2) leads to an unstable fine-tuning process.

Method

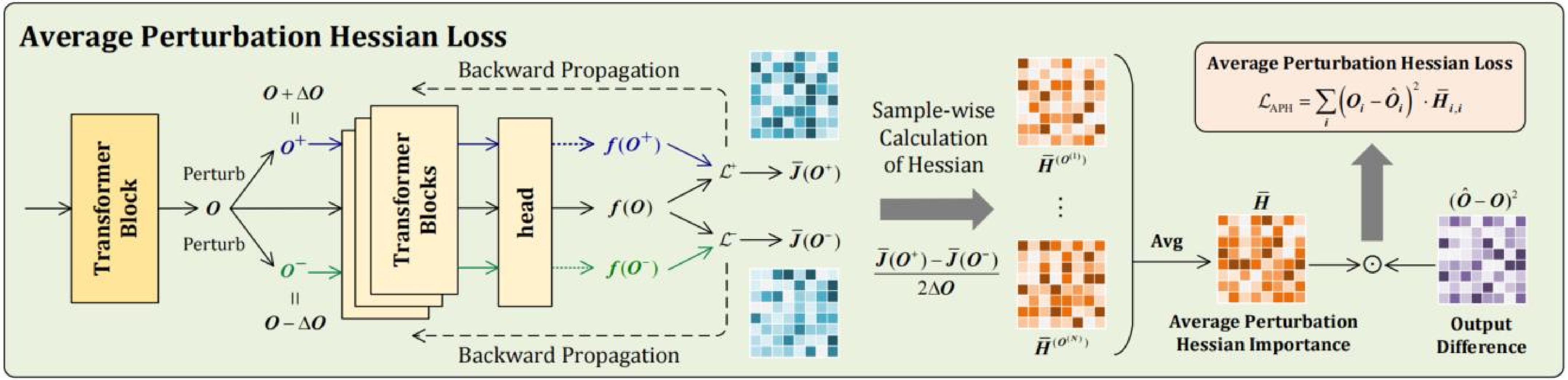
Average Perturbation Hessian Based Reconstruction (APHQ-ViT)



- **APHQ-ViT** first reconstruct the MLP layer, followed by quantization reconstruction.
- Both reconstructions are optimized by the proposed **Average Perturbation Hessian** (APH) loss.
- The **MLP Reconstruction** (MR) replaces the GELU activation with ReLU and reduces the range of post-GELU activations.

Method

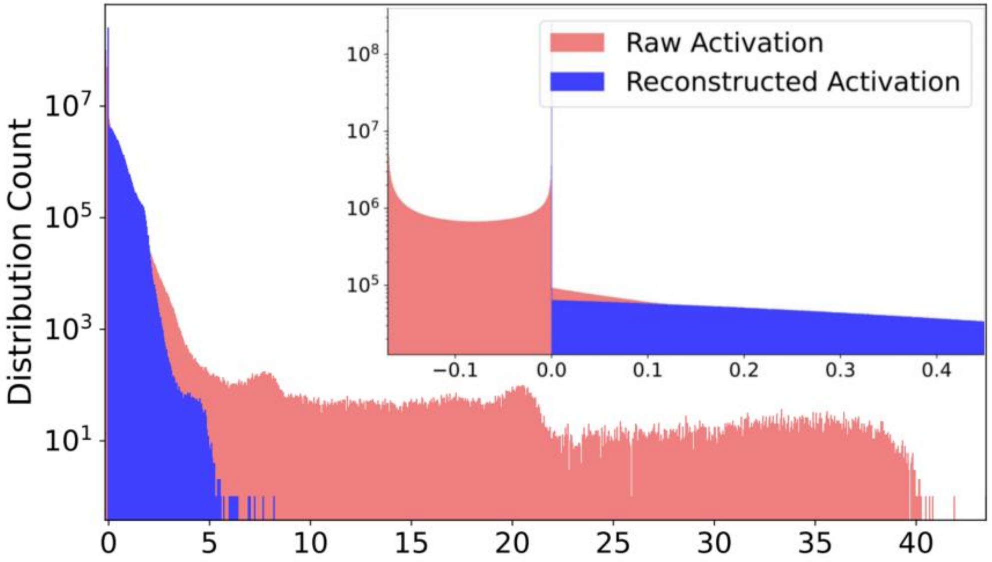
Average Perturbation Hessian Loss

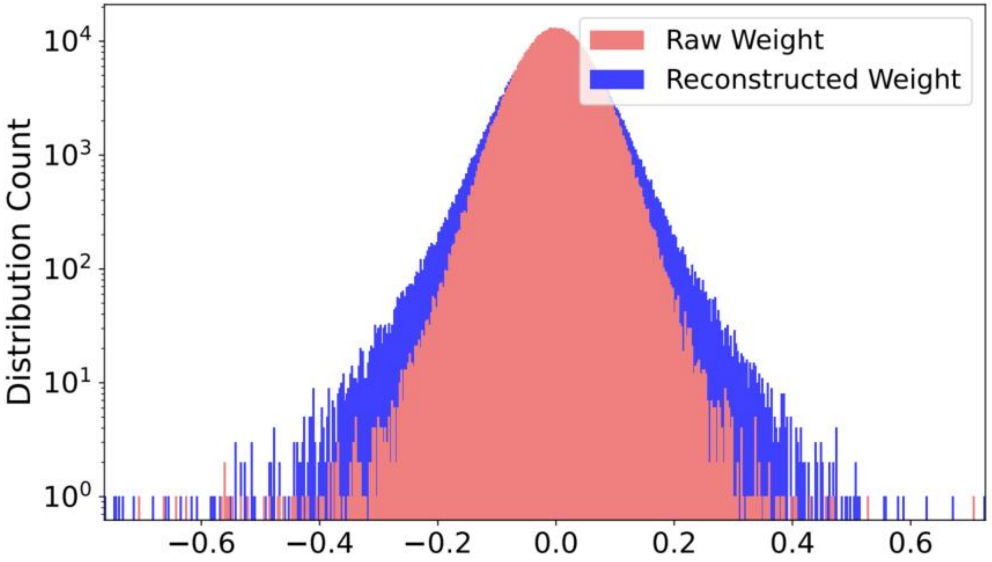


- Add a fixed perturbation $\Delta \boldsymbol{O}$ to the block output $\boldsymbol{O}$ and perform forward propagations;
- Calculate the KL divergence between the initial model output logits and the perturbed model output logits;
- Perform backward propagations to obtain the Jacobian matrices, and calculate the sample-wise Hessian;
- Calculate the Average Hessian across all samples as the output importance measurement;
- The APH loss is formulated as: $\mathcal{L}_{\mathrm{APH}} = \sum_i \left(\widehat{\boldsymbol{O}}_i - \boldsymbol{O}_i\right)^2 \cdot \bar{\boldsymbol{H}}_{i,i},$

Method

MLP Reconstruction





- We first directly replace the GELU activation with ReLU in each MLP;
- We use a clamp loss to constrain the activation range and a direct loss to prevent vanishing gradients:

$$\boldsymbol{A}_{\mathrm{FC2}} = \mathrm{ReLU}(\mathrm{FC1}(\boldsymbol{X})),$$
$$\boldsymbol{O}_{\mathrm{clamp}} = \mathrm{FC2}(\mathrm{clamp}(\boldsymbol{A}_{\mathrm{FC2}},\ \mathrm{Quantile}_p(\boldsymbol{A}_{\mathrm{FC2}}))).$$

$$\mathcal{L}_{\mathrm{Clamp}} = (\boldsymbol{O}_{\mathrm{GELU}} - \boldsymbol{O}_{\mathrm{clamp}})^2 \odot \boldsymbol{H}$$
$$\mathcal{L}_{\mathrm{Direct}} = (\boldsymbol{O}_{\mathrm{GELU}} - \boldsymbol{O}_{\mathrm{Direct}})^2 \odot \boldsymbol{H}$$

- MLP Reconstruction effectively narrows the range of post-GELU activations while preserving the weight distribution, resulting in only a minimal performance decline.

Experiments

Main Results on ImageNet

Method	Opt.	PSQ	PGQ	W/A	ViT-S	ViT-B	DeiT-T	DeiT-S	DeiT-B	Swin-S	Swin-B
Full-Prec.	-	-	-	32/32	81.39	84.54	72.21	79.85	81.80	83.23	85.27
PTQ4ViT [50]	×	TUQ	TUQ	3/3	0.10	0.10	3.50	0.10	31.06	28.69	20.13
RepQ-ViT [27]	×	$\log\sqrt{2}$	Uniform	3/3	0.10	0.10	0.10	0.10	0.10	0.10	0.10
AdaLog [47]	×	AdaLog	AdaLog	3/3	13.88	37.91	31.56	24.47	57.47	64.41	69.75
I&S-ViT [54]	✓	SULQ	Uniform	3/3	45.16	63.77	41.52	55.78	73.30	74.20	69.30
DopQ-ViT [49]	✓	TanQ	Uniform	3/3	54.72	65.76	44.71	59.26	74.91	74.77	69.63
QDrop* [46]	✓	Uniform	Uniform	3/3	38.31	73.79	46.69	52.55	74.32	74.11	75.28
APHQ-ViT(Ours)	✓	Uniform	Uniform	3/3	**63.17**	**76.31**	**55.42**	**68.76**	**76.31**	**76.10**	**78.14**
PTQ4ViT [50]	×	TUQ	TUQ	4/4	42.57	30.69	36.96	34.08	64.39	76.09	74.02
APQ-ViT [8]	×	MPQ	Uniform	4/4	47.95	41.41	47.94	43.55	67.48	77.15	76.48
RepQ-ViT [27]	×	$\log\sqrt{2}$	Uniform	4/4	65.05	68.48	57.43	69.03	75.61	79.45	78.32
ERQ [55]	×	$\log\sqrt{2}$	Uniform	4/4	68.91	76.63	60.29	72.56	78.23	80.74	82.44
IGQ-ViT [38]	×	GUQ	GUQ	4/4	73.61	79.32	62.45	74.66	79.23	80.98	83.14
AdaLog [47]	×	AdaLog	AdaLog	4/4	72.75	79.68	63.52	72.06	78.03	80.77	82.47
I&S-ViT [54]	✓	SULQ	Uniform	4/4	74.87	80.07	65.21	75.81	79.97	81.17	82.60
DopQ-ViT [49]	✓	TanQ	Uniform	4/4	75.69	80.95	65.54	75.84	80.13	81.71	83.34
QDrop* [46]	✓	Uniform	Uniform	4/4	67.62	82.02	64.95	74.73	79.64	81.03	82.79
OASQ [36]	✓	Unifrom	Uniform	4/4	72.88	76.59	66.31	76.00	78.83	81.02	82.46
APHQ-ViT(Ours)	✓	Uniform	Uniform	4/4	**76.07**	**82.41**	**66.66**	**76.40**	**80.21**	**81.81**	**83.42**

Experiments

Main Results on COCO

Method	Opt.	PSQ	W/A	Mask R-CNN				Cascade Mask R-CNN			
				Swin-T		Swin-S		Swin-T		Swin-S	
				AP^b	AP^m	AP^b	AP^m	AP^b	AP^m	AP^b	AP^m
Full-Precision	-	-	32/32	46.0	41.6	48.5	43.3	50.4	43.7	51.9	45.0
Baseline*	×	Uniform	4/4	34.6	34.2	40.8	38.6	45.9	40.2	47.9	41.6
RepQ-ViT [27]	×	$\log\sqrt{2}$	4/4	36.1	36.0	$44.2_{42.7}$†	40.2	47.0	41.1	49.3	43.1
ERQ [55]	×	$\log\sqrt{2}$	4/4	36.8	36.6	43.4	40.7	47.9	42.1	50.0	43.6
I&S-ViT [54]	✓	SULQ	4/4	37.5	36.6	43.4	40.3	48.2	42.0	**50.3**	43.6
DopQ-ViT [49]	✓	TanQ	4/4	37.5	36.5	43.5	40.4	48.2	42.1	**50.3**	**43.7**
QDrop* [46]	✓	Uniform	4/4	36.2	35.4	41.6	39.2	47.0	41.3	49.0	42.5
APHQ-ViT (Ours)	✓	Uniform	4/4	**38.9**	**38.1**	**44.1**	**41.0**	**48.9**	**42.7**	**50.3**	**43.7**

Experiments

Ablation Study

Effect of main components

Method	ViT-S	ViT-B	DeiT-T	DeiT-S	Swin-S
Full-Prec.	81.39	84.54	72.21	79.85	81.80
QDrop	38.31	73.79	46.69	52.55	74.11
+APH	59.11	76.05	53.82	67.40	75.44
+APH +MR	**63.17**	**76.31**	**55.42**	**68.76**	**76.10**

Detailed ablation of APH and MR

Method	ViT-S	ViT-B	DeiT-T	DeiT-S	Swin-S
Full-Prec.	81.39	84.54	72.21	79.85	83.23
MSE [46]	38.31	73.79	46.69	52.55	74.11
BH [25]	54.33	66.62	49.27	63.72	75.20
PH	55.14	72.80	52.25	66.12	75.40
APH	**59.11**	**76.05**	**53.82**	**67.40**	**75.44**

Method	ViT-S	ViT-B	DeiT-T	DeiT-S	Swin-S
Full-Prec.	81.39	84.54	72.21	79.85	83.23
MLP Recon.	80.90	84.84	71.07	79.38	83.12

Training Efficiency

Model	Method	PTQ	Data Size	Time Cost	Acc.
DeiT-S	LSQ [12]	×	1280 K	~170 h	77.3
	QDrop [46]	✓	1024	47 min	52.6
	APHQ-ViT	✓	1024	62 min	68.8
Swin-S	LSQ [12]	×	1280 K	~450 h	80.6
	QDrop [46]	✓	1024	130 min	74.1
	APHQ-ViT	✓	1024	170 min	76.1

Inference Efficiency

Model	AF	Bits	Lat.	TP	SR
DeiT-T	GELU	32	30.93	32.08	×1
	GELU	8	22.34	44.76	× 1.40
	ReLU	8	**20.66**	**48.40**	× **1.51**

Conclusion

IRIP laboratory

- We propose a novel **post-training quantization approach** APHQ-ViT for ViTs. Notably, compared to the state-of-the-art methods, APHQ-ViT achieves an average improvement of 7.21% on ImageNet with 3-bit quantization using only uniform quantizers.

- We demonstrate that the current Hessian guided loss adopts an inaccurate estimated Hessian matrix, and present an improved Average Perturbation Hessian (APH) loss. Based on APH, we develop an MLP Reconstruction method that simultaneously replaces the GELU activation function with ReLU and significantly reduces the activation range.

- Extensive experimental results show the effectiveness of our approach across various Vision Transformer architectures and vision tasks, including image classification, object detection, and instance segmentation.

Thanks!