Learning And Mining from DatA

Quantization without Tears

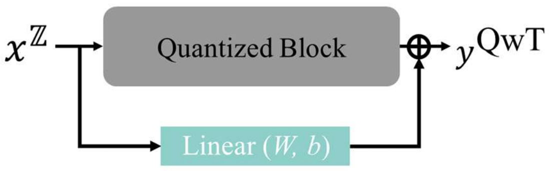
Minghao Fu, Hao Yu, Jie Shao, Junjie Zhou, Ke Zhu, Jianxin Wu
National Key Laboratory for Novel Software Technology, Nanjing University, China
School of Artificial Intelligence, Nanjing University, China

CVPR Nashville JUNE 11-15, 2025

Background & Motivation

Summary:

- QwT generates a quantized network.
- It gradually compensates for the information loss introduced during the quantization of **each block** by incorporating **full-precision linear layers**.



Advantages: 💥

- ***Speed:*** The process is completed in ~2 minutes.
- ***Simplicity:*** No tedious hyperparameter tuning. The compensation module, based on simple linear layers, has a *closed-form* solution.
- ***Generality:*** Applicable across a variety of
  - ✓ architectures—CNNs, Transformers, LLMs, DiTs;
  - ✓ tasks—Recognition, Detection, Generation.
- ***Practical Deployment:*** QwT can be integrated with existing PTQ methods and deployed on infrastructures that support fixed-point inference.

Method Overview

Key Idea:

✓ QwT uses lightweight linear layers to counteract the information loss due to quantization.

Process:

① Apply *any* quantization method to obtain the quantized model: $\{l\} \Longrightarrow \{l^{\mathbb{Z}}\}$.

② Get the quantized output: $Y^{\mathbb{Z}} = l^{\mathbb{Z}}(X^{\mathbb{Z}})$.

③ Get the FP output: $Y = l(X^{\mathbb{Z}})$.

④ Get $\{W, b\}$ using linear regression: $\{X^{\mathbb{Z}}, Y - Y^{\mathbb{Z}}\}$.

⑤ Finish compensation: $Y^{QwT} = l^{\mathbb{Z}}(X^{\mathbb{Z}})+WX^{\mathbb{Z}} + b$.

Model Size & Inference Latency

Network	Method	Size	Latency	Top-1
DeiT-T	Full-precision	22.9	11.6	72.2
	Percentile [23]	5.9	2.8	71.2
	Percentile + QwT	6.8	3.2	71.5
Swin-T	Full-precision	113.2	34.5	81.4
	Percentile [23]	28.6	9.5	80.8
	Percentile + QwT	32.9	10.9	81.0
Swin-S	Full-precision	198.4	61.0	83.2
	Percentile [23]	50.1	16.0	82.1
	Percentile + QwT	58.0	17.9	83.0
ViT-S	Full-precision	88.2	28.3	81.4
	Percentile [23]	22.5	5.8	79.2
	Percentile + QwT	26.0	6.6	80.1
ViT-B	Full-precision	346.3	85.3	84.5
	Percentile [23]	87.4	15.5	75.8
	Percentile + QwT	101.6	17.5	82.8

Main Results

Network	Method	#Bits	Size	Top-1
DeiT-T	Full-precision	32/32	22.9	72.2
	IGQ-ViT† [38]	4/4	-	62.5
	RepQ-ViT [27]	4/4	3.3	58.2
	RepQ-ViT + QwT	4/4	4.2	61.4
	RepQ-ViT + QwT*	4/4	4.2	**64.8**
	IGQ-ViT† [38]	6/6	-	71.2
	RepQ-ViT [27]	6/6	4.6	71.0
	RepQ-ViT + QwT	6/6	5.5	71.2
	RepQ-ViT + QwT*	6/6	5.5	**71.6**
Swin-T	Full-precision	32/32	113.2	81.4
	IGQ-ViT† [38]	4/4	-	77.8
	RepQ-ViT [27]	4/4	14.9	73.0
	RepQ-ViT + QwT	4/4	19.2	75.5
	RepQ-ViT + QwT*	4/4	19.2	**79.3**
	IGQ-ViT† [38]	6/6	-	80.9
	RepQ-ViT [27]	6/6	21.7	80.6
	RepQ-ViT + QwT	6/6	26.0	80.7
	RepQ-ViT + QwT*	6/6	26.0	**80.9**
ResNet-50	Full-precision	32/32	102.2	76.6
	CL-Calib† [47]	4/4	-	75.4
	Percentile[23]	4/4	14.0	68.4
	Percentile + QwT	4/4	16.0	74.5
	Percentile + QwT*	4/4	16.0	**75.8**
	CL-Calib† [47]	6/6	-	-
	Percentile[23]	6/6	19.9	76.0
	Percentile + QwT	6/6	21.9	76.8
	Percentile + QwT*	6/6	21.9	**76.8**

Image Classification

Network	Method	#Bits	Size	AP$^{\text{box}}$	AP$^{\text{mask}}$
ResNet-50 + DETR	Full-precision	32/32	164.5	42.0	-
	MinMax	6/6	47.4	39.5	-
	MinMax + QwT	6/6	49.4	**40.0**	-
	MinMax	8/8	56.4	41.6	-
	MinMax + QwT	8/8	58.4	**41.7**	-
Swin-S + Mask R-CNN	Full-precision	32/32	276.5	48.5	43.3
	RepQ-ViT [27]	4/4	36.1	42.6	40.0
	RepQ-ViT + QwT	4/4	44.0	**43.1**	**40.4**
	RepQ-ViT [27]	6/6	53.3	47.6	42.9
	RepQ-ViT + QwT	6/6	61.2	**48.0**	**43.1**
Swin-S + Cascade Mask R-CNN	Full-precision	32/32	427.8	51.9	45.0
	RepQ-ViT [27]	4/4	56.9	49.3	43.1
	RepQ-ViT + QwT	4/4	64.8	**49.9**	**43.4**
	RepQ-ViT [27]	6/6	83.4	51.4	44.6
	RepQ-ViT + QwT	6/6	91.3	**51.7**	**44.8**
Swin-B + Cascade Mask R-CNN	Full-precision	32/32	579.9	51.9	45.0
	RepQ-ViT [27]	4/4	76.1	49.3	43.1
	RepQ-ViT + QwT	4/4	90.1	**50.0**	**43.7**
	RepQ-ViT [27]	6/6	112.1	51.5	44.8
	RepQ-ViT + QwT	6/6	126.1	**51.8**	**45.0**

Detection & Segmentation

Method	#Bits	Size (GB)	W2 (↓)	C4 (↓)	QA. Avg (↑)
Full-precision	16	16.06	6.24	8.96	66.10
GPTQ	4	5.73	6.65	9.44	64.90
GPTQ + QwT	4	6.80	**6.63**	**9.38**	**65.18**

Language Generation (LLaMA3-8B)

Method	#Bits	Size (MB)	FID (↓)	IS (↑)
Full-precision	16/16	1349	5.32	236.17
RepQ-ViT	8/8	677	5.46	234.74
GPTQ	8/8	690	5.90	218.90
Q-DiT	8/8	683	5.45	236.52
Q-DiT + QwT	8/8	707	**5.35**	**236.91**
RepQ-ViT	4/8	339	319.68	2.20
GPTQ	4/8	351	9.94	166.35
Q-DiT	4/8	347	6.75	208.38
Q-DiT + QwT	4/8	361	**6.06**	**215.70**

Image Generation (DiT-XL/2)