Poster Session

ICML
International Conference
On Machine Learning

LRA-QViT: Integrating Low-Rank Approximation and Quantization for Robust and Efficient Vision Transformers

Beom Jin Kang, Nam Joon Kim, Hyun Kim*

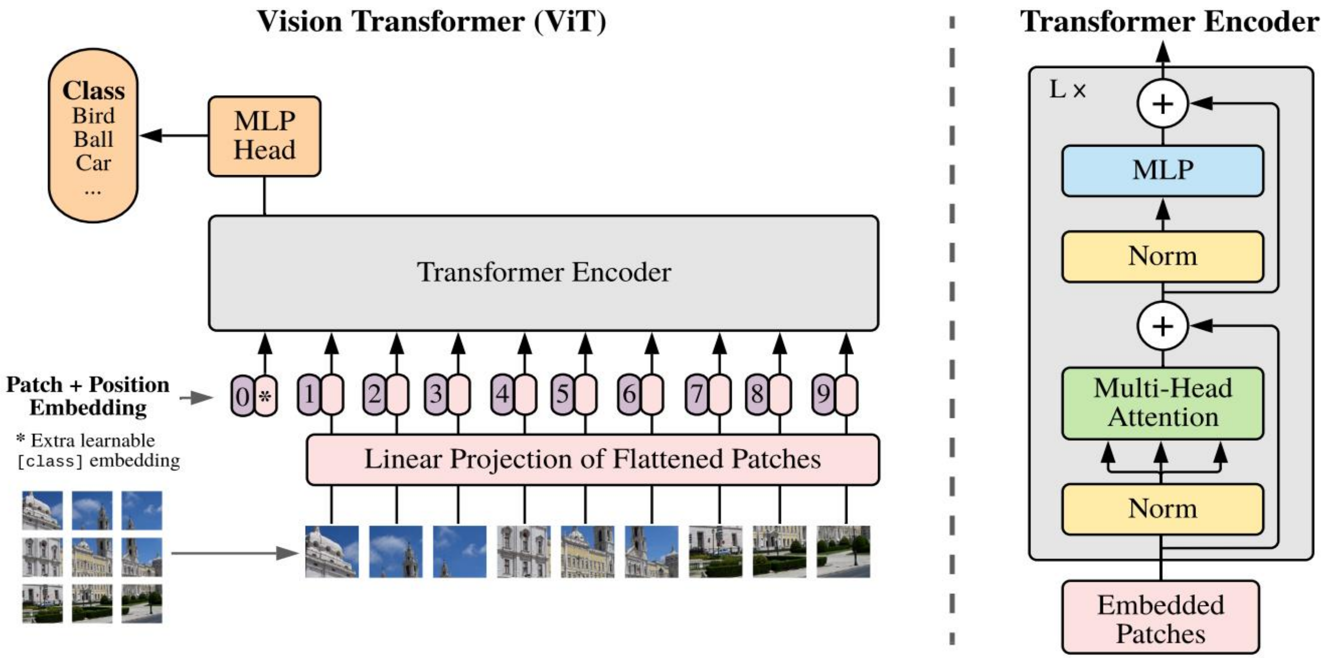
Seoul National University of Science and Technology, Seoul, Korea
{beomjin, rlarla2626, hyunkim}@seoultech.ac.kr

IDSL
Intelligent Digital Systems Design Lab.

SEOULTECH

Introduction-1

- **Vision Transformer (ViT) :**
  - **ViT Demonstrated superior accuracy in diverse Computer Vision Tasks**
  - However, **their large size and computational cost** limit their use in **resource-constrained environments** (*e.g.,* edge, mobile).
  - Consequently, various ViT compression studies have been conducted



Vision Transformer Architecture

(An image is worth 16x16 words: Transformers for image recognition at scale, ICLR, 2021)

Introduction-2

- **Motivations :**
  - Common compression techniques, such as **Low-Rank Approximation** and **Quantization**, have been explored
  - **Low-Rank Approximation (LRA) :**
    - **Singular value decomposition-based FC layer compression methods**
    - LRA studies performed knowledge-distillation based fine-tuning to recover accuracy
    - **However, reducing the fundamental information loss in the weight matrix could enable even higher accuracy**
  - **Quantization :**
    - **Compressing FP32 model weights and activations by quantizing them to lower bit-precision**
    - **When used in conjunction with LRA, it has the potential to achieve greater model size reduction than a single method**
    - However, to date, there have been no attempts to simultaneously apply both LRA and Quantization
    - **Integrating both methods requires developing quantization techniques highly compatible with LRA**

Introduction-3

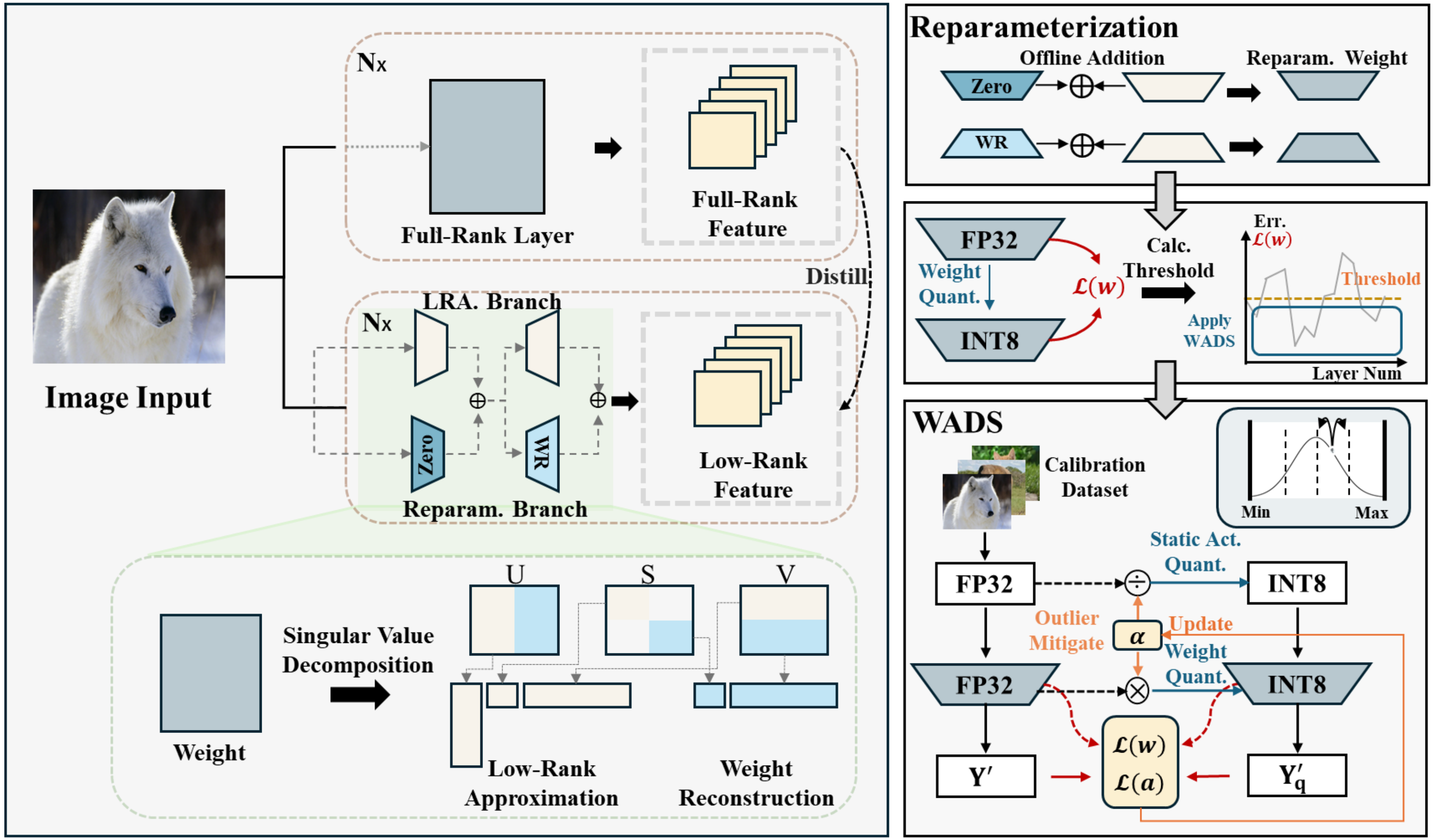
- **Goals :**
  - **Inference Efficiency** : Effectively combining LRA and quantization to achieve **higher compression ratios** and **faster inference speed** than previous single-method approaches
  - **High Accuracy : Minimizing accuracy degradation** when **applying LRA** and addressing **outlier issues** in combination **with quantization to achieve high accuracy**

◈ **Approach**

- **Developing a robust LRA method** : Proposing a low-cost error compensation matrix design and an initialization method to reduce weight information loss
- **Block-Level Knowledge Distillation:** Achieving superior generalization performance using encoder block-level knowledge distillation
- **LRA-Aware Quantization :** Proposing a distribution scaling method to minimize outlier effects when applying LRA.
- **Ultimately, combining LRA and quantization to achieve a high model compression ratio and low inference latency with minimal accuracy degradation**

Ours

- **LRA-QViT : Proposed ViT Compression Framework**



(a) RB-LRA & WR (b) Quantization

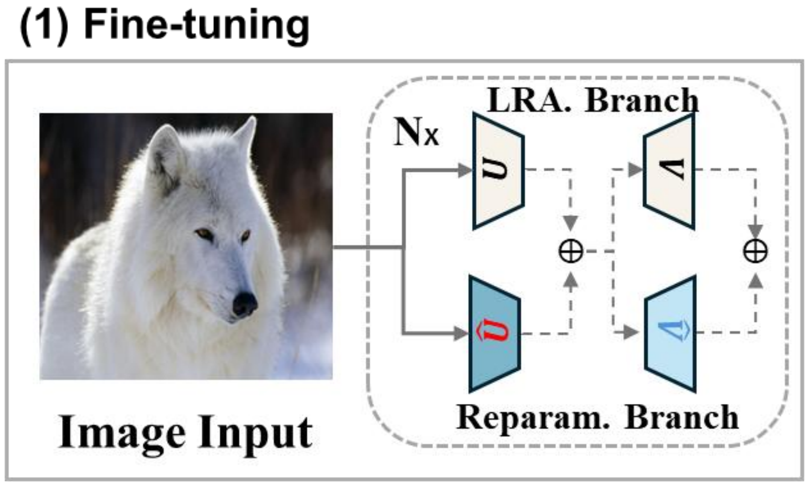
Ours

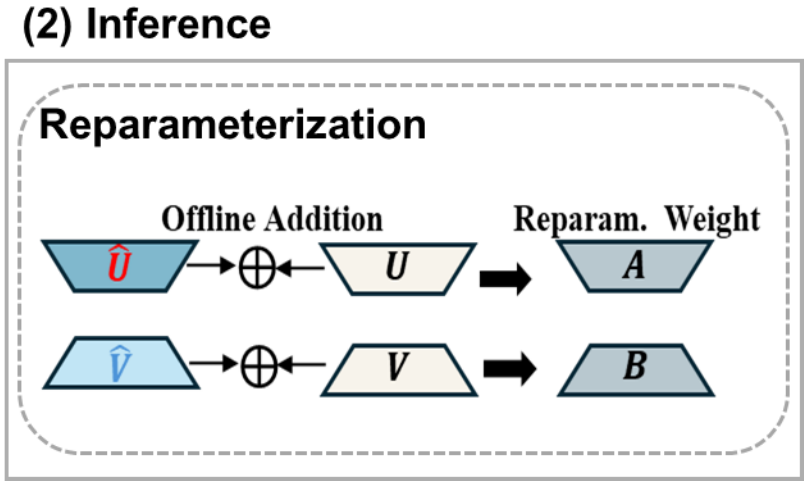
- **RB-LRA : Reparameterizable branch-based low-rank approximation**
  - Design of FC Layers with a **reparameterizable addition branch** in the form of a **low-rank matrix to compensate for LRA errors**

$$y \approx V'\left(U'^{T}x\right) + \hat{E}x = \left(V' + \hat{V}\right)\left(U'^{T} + \hat{U}^{T}\right)x \qquad (1)$$

$$where\ \hat{E}x = \left(V'\hat{U}^{T} + \hat{V}U'^{T} + \hat{V}\hat{U}^{T}\right)x$$

  - **Optimize the $\hat{V}$ and $\hat{U}$** matrices through fine-tuning
  - Apply reparameterization during inference → integrate into a single branch
  - **Reduction in parameter and computational cost**





<**RB-LRA** : **R**eparameterizable **B**ranch-based **L**ow-**R**ank **A**pproximation>

LRA-QViT : Reparameterizable branch-based LRA (RB-LRA)

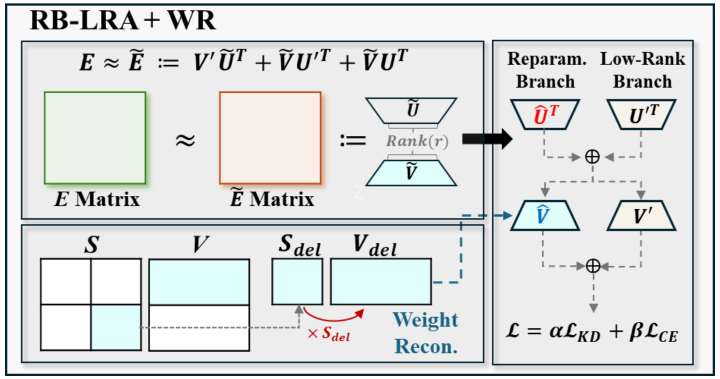
- **Weight Reconstruction**
  - **Initialize RB-LRA using the LRA removal matrix.**
    - **Matrix Removed During LRA :**

$$\begin{cases} U_{\nabla} = U_{[:,r:]} \\ S_{\nabla} = S_{[r:,r:]} \\ V_{\nabla} = V_{[r:,:]} \end{cases} \tag{2}$$

    - **Original FC Layer Reconstruction with Removed Matrix**

$$\begin{aligned} y' &= U^T x = Concat(U'^T x, U^T{}_{\nabla} x) \\ y &= VS^T y' = V' y'_{[:r,:]} + (S_{\nabla} V_{\nabla})^T y'_{[r:,:]} \end{aligned} \tag{3}$$

    - **U_{∇} matrix :** Reconstruction by **Concat** → **Residual Branch Reconstruction: Impossible** → $\hat{U}$ **: zero**
    - **$S_{\nabla} V_{\nabla}$matrix :** Reconstruction by **Addition** → **Residual Branch Reconstruction: Possible**→ $\hat{V}$ **: $S_{\nabla} V_{\nabla}$**



<**W**eight **R**econstruction>

LRA-QViT : Weight-Aware Distribution Scaling (WADS)

- **WASD : Weight-aware distribution scaling**
  - **Calculate Weight Quantization Error**
    - Scaling Applied Exclusively to Layers Below Threshold

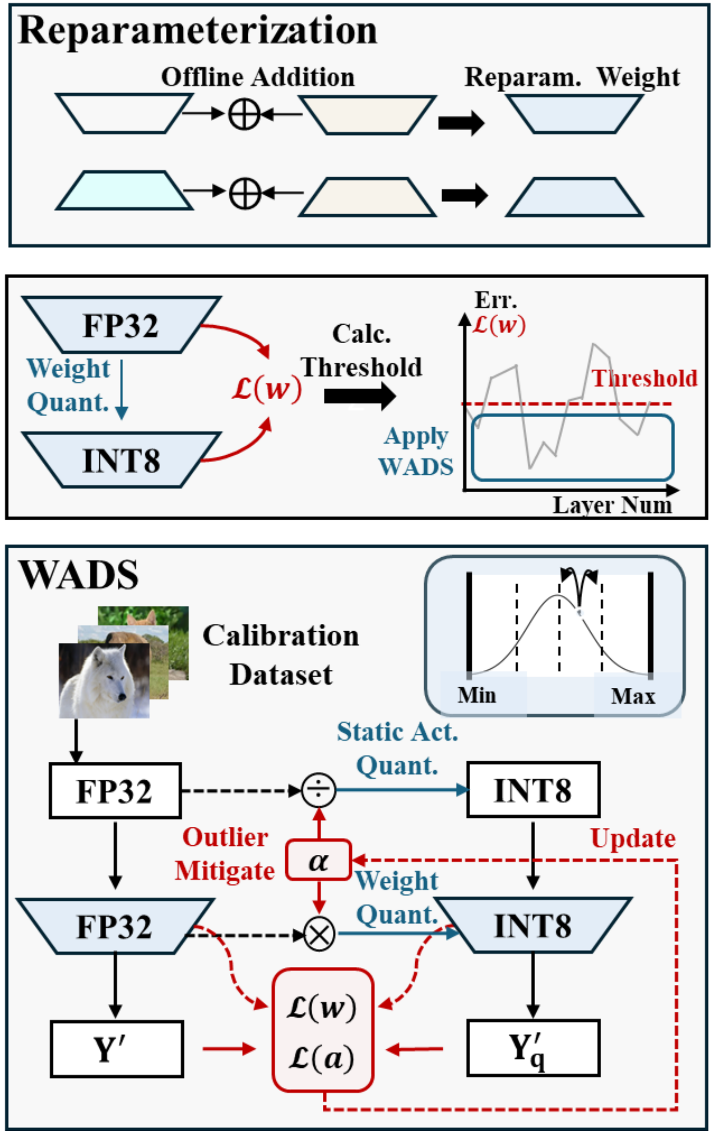
$$\mathcal{L}(w) = \|Q(w) - w\|^2 \tag{4}$$

  - **Optimal Scaling Vector Search**
    - Weight Quantization Error-Aware Loss Function Design
    - Achieving Optimal Accuracy

$$a' = \underset{\alpha}{\operatorname{argmin}} \left\| Q\left(\frac{x}{\alpha}\right) Q(\alpha w) - xw \right\|^2 + \|Q(w) - w\|^2 \tag{5}$$

  - **WADS-based Quantization**
    - s = quantization scaling factor

$$Y_q = Q\left(\frac{x}{\alpha}\right) Q(\alpha w)$$
$$Q(x) = clip\left(round\left(\frac{x}{s}\right), -2^{b-1}, 2^{b-1} - 1\right) \tag{6}$$



<**W**eight-**A**ware **D**istribution **S**caling>

Experimental Results-1 (RB-LRA)

- **Evaluation of RB-LRA : ImageNet**
  - **RB-LRA : Params → -45.7% Accuracy Drop → 0.73% (DeiT-B)**
    - **Other Models :** Achieved SOTA Accuracy
- **Other Applications**
  - **Object Detection / Instance Segmentation :**
    - **Params → –7.1M AP Drop → -0.3%**
  - **Pose Estimation :**
    - **Params → -25.7% AP / AR Drop → -0.9%**
  - **Language Processing**
    - **Params → -29.7% PPL → +0.7%**
  - **Speech Recognition**
    - **Params → -26.3% WER → +0.2%**

Model	Method	KD Method	Params(M)	GFLOPs	ACC.(%)	Diff.(%)
DeiT-T	Baseline	-	5.7	2.2	72.17	-
	LRA	-	5.2 (-8.8%)	1.8	68.40	-3.77
	RB-LRA	-	**5.2 (-8.8%)**	**1.8**	**70.92**	**-1.25**
	RB-LRA + KD	**Feature**	**5.2 (-8.8%)**	**1.8**	**71.70**	**-0.47**
DeiT-B	Baseline	-	86.6	33.7	81.85	-
	LRA	-	44.4 (-45.7%)	17.1	78.76	-3.09
	PELA (Guo et al., 2024)	Feature	44.1 (-49.1%)	17.0	81.00	-0.85
	RB-LRA	-	**44.4 (-45.7%)**	**17.1**	**79.93**	**-1.92**
	RB-LRA + KD	**Feature**	**44.4 (-45.7%)**	**17.1**	**81.12**	**-0.73**
Swin-T	Baseline	-	28.3	8.6	81.37	-
	LRA	-	21.1 (-25.4%)	6.7	77.30	-4.07
	RB-LRA	-	**21.1 (-25.4%)**	**6.7**	**80.27**	**-1.1**
	RB-LRA + KD	**Feature**	**21.1 (-25.4%)**	**6.7**	**80.49**	**-0.88**
Swin-B	Baseline	-	88.1	30.3	83.47	-
	LRA	-	60.1 (-31.8%)	21.1	81.75	-1.72
	AAFM+GFM (Yu & Wu, 2023)	Feature	60.2 (-31.7%)	-	82.99	-0.48
	PELA (Guo et al., 2024)	Feature	62.2 (-29.4%)	21.3	82.50	-0.97
	RB-LRA	-	**60.1 (-31.8%)**	**21.1**	**82.88**	**-0.59**
	RB-LRA+KD	**Feature**	**60.1 (-31.8%)**	**21.1**	**83.44**	**-0.03**

Backbone	Params(M)	GFLOPs	AP^{box}	AP^{mask}
ResNet-50 (He et al., 2016)	44.4	250.2	40.0	36.1
PVT-M (Wang et al., 2021)	63.9	351.2	42.0	39.0
Swin-T (Liu et al., 2021)	47.8	256.8	42.7	39.3
Swin-T + RB-LRA	**40.6**	**241.2**	**42.5**	**39.0**

Model	Method	Params(M)	PPL	WER
GPT-2 Medium	Baseline	354.8	18.72	-
	RB-LRA	**249.4 (-29.7%)**	**19.51**	-
Conformer-L	Baseline	116.8	-	5.4
	RB-LRA	**86.2(-26.3%)**	-	**5.6**

Model	Method	Params(M)	AP	AR
ViTPose-B	Baseline	89.9	75.9	81.0
	RB-LRA	**66.8 (-25.7%)**	**75.0**	**80.4**

Experimental Results-1 (RB-LRA)

- **Evaluation of RB-LRA : ImageNet**
  - RB-LRA : Params → -45.7% Accuracy Drop → 0.73% (DeiT-B)
    - **Other Models :** Achieved SOTA Accuracy
- **Other Applications**
  - **Object Detection / Instance Segmentation :**
    - Params → –7.1M AP Drop → -0.3%
  - **Pose Estimation :**
    - Params → -25.7% AP / AR Drop → -0.9%
  - **Language Processing**
    - Params → -29.7% PPL → +0.7%
  - **Speech Recognition**
    - Params → -26.3% WER → +0.2%

Model	Method	Params(M)	PPL	WER
GPT-2 Medium	Baseline	354.8	18.72	-
	RB-LRA	**249.4 (-29.7%)**	**19.51**	-
Conformer-L	Baseline	116.8	-	5.4
	RB-LRA	**86.2(-26.3%)**	-	**5.6**

Model	Method	KD Method	Params(M)	GFLOPs	ACC.(%)	Diff.(%)
DeiT-T	Baseline	-	5.7	2.2	72.17	-
	LRA	-	5.2 (-8.8%)	1.8	68.40	-3.77
	RB-LRA	-	**5.2 (-8.8%)**	**1.8**	**70.92**	**-1.25**
	RB-LRA + KD	**Feature**	**5.2 (-8.8%)**	**1.8**	**71.70**	**-0.47**
DeiT-B	Baseline	-	86.6	33.7	81.85	-
	LRA	-	44.4 (-45.7%)	17.1	78.76	-3.09
	PELA (Guo et al., 2024)	Feature	44.1 (-49.1%)	17.0	81.00	-0.85
	RB-LRA	-	**44.4 (-45.7%)**	**17.1**	**79.93**	**-1.92**
	RB-LRA + KD	**Feature**	**44.4 (-45.7%)**	**17.1**	**81.12**	**-0.73**
Swin-T	Baseline	-	28.3	8.6	81.37	-
	LRA	-	21.1 (-25.4%)	6.7	77.30	-4.07
	RB-LRA	-	**21.1 (-25.4%)**	**6.7**	**80.27**	**-1.1**
	RB-LRA + KD	**Feature**	**21.1 (-25.4%)**	**6.7**	**80.49**	**-0.88**
Swin-B	Baseline	-	88.1	30.3	83.47	-
	LRA	-	60.1 (-31.8%)	21.1	81.75	-1.72
	AAFM+GFM (Yu & Wu, 2023)	Feature	60.2 (-31.7%)	-	82.99	-0.48
	PELA (Guo et al., 2024)	Feature	62.2 (-29.4%)	21.3	82.50	-0.97
	RB-LRA	-	**60.1 (-31.8%)**	**21.1**	**82.88**	**-0.59**
	RB-LRA+KD	**Feature**	**60.1 (-31.8%)**	**21.1**	**83.44**	**-0.03**

Backbone	Params(M)	GFLOPs	AP^{box}	AP^{mask}
ResNet-50 (He et al., 2016)	44.4	250.2	40.0	36.1
PVT-M (Wang et al., 2021)	63.9	351.2	42.0	39.0
Swin-T (Liu et al., 2021)	47.8	256.8	42.7	39.3
Swin-T + RB-LRA	**40.6**	**241.2**	**42.5**	**39.0**

Model	Method	Params(M)	AP	AR
ViTPose-B	Baseline	89.9	75.9	81.0
	RB-LRA	**66.8 (-25.7%)**	**75.0**	**80.4**

Experimental Results-1 (RB-LRA)

- **Evaluation of RB-LRA : ImageNet**
  - **RB-LRA : Params → -45.7% Accuracy Drop → 0.73% (DeiT-B)**
    - **Other Models :** Achieved SOTA Accuracy
- **Other Applications**
  - **Object Detection / Instance Segmentation :**
    - **Params → –7.1M AP Drop → -0.3%**
  - **Pose Estimation :**
    - **Params → -25.7% AP / AR Drop → -0.9%**
  - **Language Processing**
    - **Params → -29.7% PPL → +0.7%**
  - **Speech Recognition**
    - **Params → -26.3% WER → +0.2%**

Model	Method	Params(M)	PPL	WER
GPT-2 Medium	Baseline	354.8	18.72	-
	RB-LRA	**249.4 (-29.7%)**	**19.51**	-
Conformer-L	Baseline	116.8	-	5.4
	RB-LRA	**86.2(-26.3%)**	-	**5.6**

Model	Method	KD Method	Params(M)	GFLOPs	ACC.(%)	Diff.(%)
DeiT-T	Baseline	-	5.7	2.2	72.17	-
	LRA	-	5.2 (-8.8%)	1.8	68.40	-3.77
	RB-LRA	-	**5.2 (-8.8%)**	**1.8**	**70.92**	**-1.25**
	RB-LRA + KD	**Feature**	**5.2 (-8.8%)**	**1.8**	**71.70**	**-0.47**
DeiT-B	Baseline	-	86.6	33.7	81.85	-
	LRA	-	44.4 (-45.7%)	17.1	78.76	-3.09
	PELA (Guo et al., 2024)	Feature	44.1 (-49.1%)	17.0	81.00	-0.85
	RB-LRA	-	**44.4 (-45.7%)**	**17.1**	**79.93**	**-1.92**
	RB-LRA + KD	**Feature**	**44.4 (-45.7%)**	**17.1**	**81.12**	**-0.73**
Swin-T	Baseline	-	28.3	8.6	81.37	-
	LRA	-	21.1 (-25.4%)	6.7	77.30	-4.07
	RB-LRA	-	**21.1 (-25.4%)**	**6.7**	**80.27**	**-1.1**
	RB-LRA + KD	**Feature**	**21.1 (-25.4%)**	**6.7**	**80.49**	**-0.88**
Swin-B	Baseline	-	88.1	30.3	83.47	-
	LRA	-	60.1 (-31.8%)	21.1	81.75	-1.72
	AAFM+GFM (Yu & Wu, 2023)	Feature	60.2 (-31.7%)	-	82.99	-0.48
	PELA (Guo et al., 2024)	Feature	62.2 (-29.4%)	21.3	82.50	-0.97
	RB-LRA	-	**60.1 (-31.8%)**	**21.1**	**82.88**	**-0.59**
	RB-LRA+KD	**Feature**	**60.1 (-31.8%)**	**21.1**	**83.44**	**-0.03**

Backbone	Params(M)	GFLOPs	AP^{box}	AP^{mask}
ResNet-50 (He et al., 2016)	44.4	250.2	40.0	36.1
PVT-M (Wang et al., 2021)	63.9	351.2	42.0	39.0
Swin-T (Liu et al., 2021)	47.8	256.8	42.7	39.3
Swin-T + RB-LRA	**40.6**	**241.2**	**42.5**	**39.0**

Model	Method	Params(M)	AP	AR
ViTPose-B	Baseline	89.9	75.9	81.0
	RB-LRA	**66.8 (-25.7%)**	**75.0**	**80.4**

Experimental Results-2 (RB-LRA + WADS)

- **Evaluation WADS : ImageNet**
  - Baseline : RB-LRA
  - **Achieving the highest accuracy**
  - Demonstrating excellent compatibility with proposed RB-LRA
  - **Superiority of the Unified Framework : Model Size Reduction: Up to 87.2%**
- **Inference Latency on Real Devices**
  - Android : Cortex-X3
  - Edge : NVIDIA Jetson AGX Xavier
  - **RB-LRA: Up to 2.1x Acceleration**
  - **RB-LRA + WADS: Up to 3.2x Acceleration**
  - **Demonstrating On-Device Acceleration of the Proposed RB-LRA + WADS Framework**

Model	Method	Prec.	Size(MB)	ACC.(%)	Diff.(%)
DeiT-T	Baseline(RB-LRA)	FP32	20.8	71.70	-
	NaivePTQ	INT8	5.2	70.90	-0.80
	SmoothQuant (Xiao et al., 2023)			71.43	-0.27
	Repq-ViT (Li et al., 2023)			71.38	-0.32
	QADS (Kim et al., 2024)			71.40	-0.30
	WADS			**71.52**	**-0.18**
DeiT-B	Baseline(RB-LRA)	FP32	177.6	81.12	-
	NaivePTQ	INT8	44.4	79.62	-1.50
	SmoothQuant (Xiao et al., 2023)			80.26	-0.86
	Repq-ViT (Li et al., 2023)			80.37	-0.75
	QADS (Kim et al., 2024)			79.82	-1.30
	WADS			**80.56**	**-0.56**
Swin-T	Baseline(RB-LRA)	FP32	84.4	80.49	-
	NaivePTQ	INT8	21.1	78.30	-2.19
	SmoothQuant (Xiao et al., 2023)			80.00	-0.49
	Repq-ViT (Li et al., 2023)			80.08	-0.41
	QADS (Kim et al., 2024)			80.04	-0.45
	WADS			**80.20**	**-0.29**
Swin-B	Baseline(RB-LRA)	FP32	240.4	83.44	-
	NaivePTQ	INT8	60.1	82.14	-1.30
	SmoothQuant (Xiao et al., 2023)			82.76	-0.68
	QADS (Kim et al., 2024)			82.37	-1.07
	WADS			**82.97**	**-0.47**

Model	Method	Prec.	Size(MB)	Android(ms)	Xavier(ms)
DeiT-B	Baseline	FP32	346.4	275.6	150.7
	RB-LRA	FP32	177.6	153.2	73.6
	RB-LRA + WADS	INT8	44.4	86.7	59.4
Swin-T	Baseline	FP32	113.2	98.5	61.1
	RB-LRA	FP32	84.4	83.6	38.6
	RB-LRA + WADS	INT8	21.1	67.3	27.4
Swin-B	Baseline	FP32	352.4	287.4	140.5
	RB-LRA	FP32	240.4	226.3	102.2
	RB-LRA + WADS	INT8	60.1	155.3	96.2

Experimental Results-2 (RB-LRA + WADS)

- **Evaluation WADS : ImageNet**
  - Baseline : RB-LRA
  - **Achieving the highest accuracy**
  - Demonstrating excellent compatibility with proposed RB-LRA
  - **Superiority of the Unified Framework :**
    Model Size Reduction: Up to 87.2%
- **Inference Latency on Real Devices**
  - Android : Cortex-X3
  - Edge : NVIDIA Jetson AGX Xavier
  - **RB-LRA: Up to 2.1x Acceleration**
  - **RB-LRA + WADS: Up to 3.2x Acceleration**
  - **Demonstrating On-Device Acceleration of the Proposed RB-LRA + WADS Framework**

Model	Method	Prec.	Size(MB)	ACC.(%)	Diff.(%)
DeiT-T	Baseline(RB-LRA)	FP32	20.8	71.70	-
	NaivePTQ	INT8	5.2	70.90	-0.80
	SmoothQuant (Xiao et al., 2023)			71.43	-0.27
	Repq-ViT (Li et al., 2023)			71.38	-0.32
	QADS (Kim et al., 2024)			71.40	-0.30
	WADS			**71.52**	**-0.18**
DeiT-B	Baseline(RB-LRA)	FP32	177.6	81.12	-
	NaivePTQ	INT8	44.4	79.62	-1.50
	SmoothQuant (Xiao et al., 2023)			80.26	-0.86
	Repq-ViT (Li et al., 2023)			80.37	-0.75
	QADS (Kim et al., 2024)			79.82	-1.30
	WADS			**80.56**	**-0.56**
Swin-T	Baseline(RB-LRA)	FP32	84.4	80.49	-
	NaivePTQ	INT8	21.1	78.30	-2.19
	SmoothQuant (Xiao et al., 2023)			80.00	-0.49
	Repq-ViT (Li et al., 2023)			80.08	-0.41
	QADS (Kim et al., 2024)			80.04	-0.45
	WADS			**80.20**	**-0.29**
Swin-B	Baseline(RB-LRA)	FP32	240.4	83.44	-
	NaivePTQ	INT8	60.1	82.14	-1.30
	SmoothQuant (Xiao et al., 2023)			82.76	-0.68
	QADS (Kim et al., 2024)			82.37	-1.07
	WADS			**82.97**	**-0.47**

Model	Method	Prec.	Size(MB)	Android(ms)	Xavier(ms)
DeiT-B	Baseline	FP32	346.4	275.6	150.7
	RB-LRA	FP32	177.6	153.2	73.6
	RB-LRA + WADS	INT8	44.4	86.7	59.4
Swin-T	Baseline	FP32	113.2	98.5	61.1
	RB-LRA	FP32	84.4	83.6	38.6
	RB-LRA + WADS	INT8	21.1	67.3	27.4
Swin-B	Baseline	FP32	352.4	287.4	140.5
	RB-LRA	FP32	240.4	226.3	102.2
	RB-LRA + WADS	INT8	60.1	155.3	96.2

Experimental Results-2 (RB-LRA + WADS)

- **Evaluation WADS : ImageNet**
  - Baseline : RB-LRA
  - **Achieving the highest accuracy**
  - Demonstrating excellent compatibility with proposed RB-LRA
  - **Superiority of the Unified Framework : Model Size Reduction: Up to 87.2%**
- **Inference Latency on Real Devices**
  - Android : Cortex-X3
  - Edge : NVIDIA Jetson AGX Xavier
  - **RB-LRA: Up to 2.1x Acceleration**
  - **RB-LRA + WADS: Up to 3.2x Acceleration**
  - **Demonstrating On-Device Acceleration of the Proposed RB-LRA + WADS Framework**

Model	Method	Prec.	Size(MB)	ACC.(%)	Diff.(%)
DeiT-T	Baseline(RB-LRA)	FP32	20.8	71.70	-
	NaivePTQ	INT8	5.2	70.90	-0.80
	SmoothQuant (Xiao et al., 2023)			71.43	-0.27
	Repq-ViT (Li et al., 2023)			71.38	-0.32
	QADS (Kim et al., 2024)			71.40	-0.30
	WADS			**71.52**	**-0.18**
DeiT-B	Baseline(RB-LRA)	FP32	177.6	81.12	-
	NaivePTQ	INT8	44.4	79.62	-1.50
	SmoothQuant (Xiao et al., 2023)			80.26	-0.86
	Repq-ViT (Li et al., 2023)			80.37	-0.75
	QADS (Kim et al., 2024)			79.82	-1.30
	WADS			**80.56**	**-0.56**
Swin-T	Baseline(RB-LRA)	FP32	84.4	80.49	-
	NaivePTQ	INT8	21.1	78.30	-2.19
	SmoothQuant (Xiao et al., 2023)			80.00	-0.49
	Repq-ViT (Li et al., 2023)			80.08	-0.41
	QADS (Kim et al., 2024)			80.04	-0.45
	WADS			**80.20**	**-0.29**
Swin-B	Baseline(RB-LRA)	FP32	240.4	83.44	-
	NaivePTQ	INT8	60.1	82.14	-1.30
	SmoothQuant (Xiao et al., 2023)			82.76	-0.68
	QADS (Kim et al., 2024)			82.37	-1.07
	WADS			**82.97**	**-0.47**

Model	Method	Prec.	Size(MB)	Android(ms)	Xavier(ms)
DeiT-B	Baseline	FP32	346.4	275.6	150.7
	RB-LRA	FP32	177.6	153.2	73.6
	RB-LRA + WADS	INT8	44.4	86.7	59.4
Swin-T	Baseline	FP32	113.2	98.5	61.1
	RB-LRA	FP32	84.4	83.6	38.6
	RB-LRA + WADS	INT8	21.1	67.3	27.4
Swin-B	Baseline	FP32	352.4	287.4	140.5
	RB-LRA	FP32	240.4	226.3	102.2
	RB-LRA + WADS	INT8	60.1	155.3	96.2

Thank you!