LampQ: Towards Accurate Layer-wise Mixed Precision Quantization for Vision Transformers

Minjun Kim, Jaeri Lee, Jongjin Kim, Jeongin Yun, Yongmo Kwon, and U Kang*

{minjun.kim, jlunits2, j2kim99, yji00828, rnjsdydah, ukang}@snu.ac.kr, *: Corresponding Author

Seoul National University, Seoul, South Korea

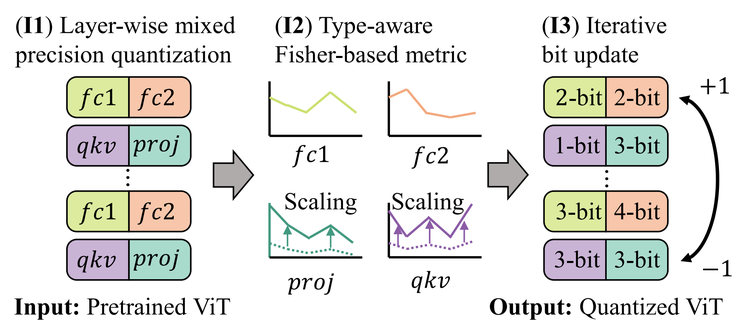
Paper

GitHub

Summary

LampQ (Layer-wise Mixed Precision Quantization for Vision Transformers)

- **TL;DR**: LampQ enables accurate layer-wise mixed precision quantization for ViTs through 1) layer-wise granularity, 2) type-aware Fisher metric, and 3) iterative bit assignment
- **GitHub**: https://github.com/snudatalab/LampQ



Problem Definition

Problem. Post-training Quantization for ViTs

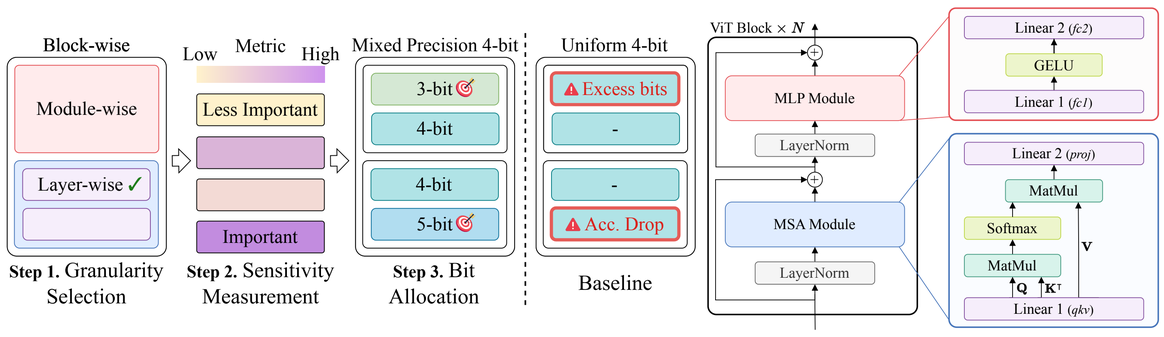
- **Input**: a ViT model f_θ pre-trained on task $\mathcal{T}$, calibration dataset $\mathbb{D} = \{(x_i, y_i)\}_{i=1}^{S}$ of size S, and target bit-width b
- **Output**: a quantized model $f_{\theta'}$ within the b-bit limit minimizing the performance degradation on $\mathcal{T}$

Mixed Precision Quantization (MPQ)

- Assigning different quantization bit-widths to each component
- **Why?** Aims to improve performance by allocating higher bit-widths to crucial components
- **Four solutions:** 1) Learning-based, 2) Reinforcement Learning (RL), 3) Neural Architecture Search (NAS), and **4) Metric-based**

Metric-base MPQ

- Bit-width allocation based on sensitivity metrics
- Most practical and scalable solution (others require training)
- **3 components: 1) granularity, 2) metric, and 3) bit assignment**



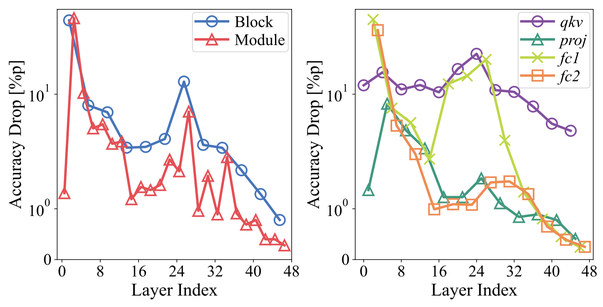
Simplified Formulation of ViT Layers

- Block > Module (MLP and MSA) > Layer ($qkv, proj, fc1$, and $fc2$)
- Abstracting each ViT block as consisting of 4 core layers
  ➔ omitting the inner workings of the Attention mechanism

Proposed Method

Granularity: Layer-wise MPQ (I1)

- **Challenge.** Coarse-grained granularity ignores the difference in layer-wise sensitivity
- **Idea.** Allocate different bit-widths for each layer to reflect their sensitivity



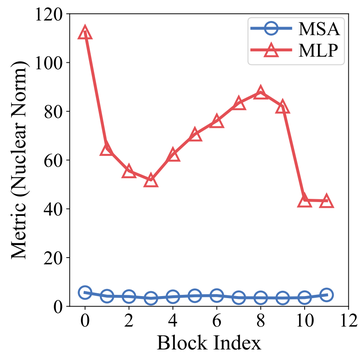
(a) Block and module (b) Layer

Metric: Type-aware Fisher-based metric (I2)

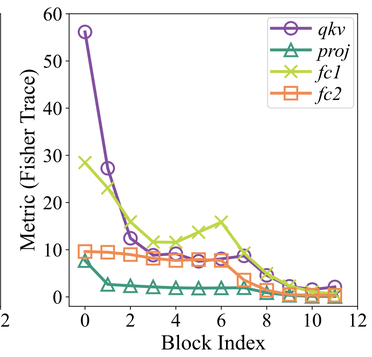
- **Challenge.** Mismatch in metric scale across component types
- **Idea.** Type-scaled sensitivity metric Ω_i:

$$\Omega_i = \alpha_t \mathrm{tr}(\mathbf{F}_i)$$

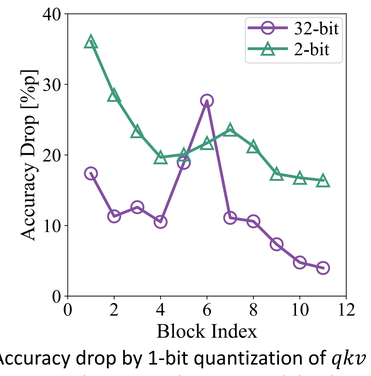
- $\alpha_t \in \{\alpha_{qkv}, \alpha_{proj}, \alpha_{fc1}, \alpha_{fc2}\}$ assigned based on layer type t



(a) VT-PTQ (2021b)



(b) LampQ (Proposed)



Accuracy drop by 1-bit quantization of qkv layers; 32-bit: original DeiT-S model, 2-bit: its variant with the first block quantized to 2-bit

Bit assignment: Iterative bit update (I3)

- **Challenge.** Sensitivity trend changes after quantization

Step 1. Initial assignment

- Initialize by solving an Integer Linear Programming (ILP) task

Step 2. Error-based iterative updates

- Iteratively balance the bit-widths while ensuring that the average bit-width is preserved (efficient computation by estimating the reconstruction error with its expected value)

Experiments

1. LampQ achieves the state-of-the-art performance

- Regardless of model type, dataset, target bit-width b, and task $\mathcal{T}$

Method	W/A	ViT		DeiT			Swin		Average
		ViT-S	ViT-B	DeiT-T	DeiT-S	DeiT-B	Swin-S	Swin-B	
Full-Precision	32/32	81.38	84.53	72.13	79.83	81.80	83.23	85.27	81.17
RepQ-ViT (2023)	4/4	65.05	68.48	57.43	69.03	75.61	79.45	78.32	70.48
OPTQ (2023)	4/4	67.59	75.12	58.96	70.85	76.10	80.17	81.08	72.84
ERQ (2024)	4/4	68.91	76.63	60.29	72.56	78.23	80.74	82.44	74.26
AdaLog (2024)	4/4	72.75	79.68	63.52	72.06	78.03	80.77	82.47	75.61
VT-PTQ† (2021b)	4MP/4MP	73.69	80.10	63.90	72.78	78.30	80.96	82.80	76.07
LampQ (Proposed)	4MP/4MP	**74.02**	**81.91**	**65.71**	**75.40**	**79.24**	**81.76**	**83.87**	**77.42**
RepQ-ViT (2023)	3/3	0.43	0.14	0.97	4.37	4.84	8.84	1.34	2.99
AdaLog (2024)	3/3	13.88	37.91	31.56	24.47	57.45	64.41	69.75	42.78
VT-PTQ† (2021b)	3MP/3MP	16.62	42.13	32.98	26.37	60.14	69.80	73.51	45.94
LampQ (Proposed)	3MP/3MP	**23.06**	**48.53**	**37.54**	**45.38**	**61.44**	**70.91**	**75.82**	**51.81**

†: AdaLog quantization with bit allocation by VT-PTQ.

- **Tasks:** Image classification, Object detection, Zero-shot quantization
- Up to **5.87%p** (3-bit)

Method	W/A	DeiT		Swin	
		DeiT-T	DeiT-S	Swin-T	Swin-S
Original	32/32	72.21	79.85	81.35	83.20
PSAQ-ViT	4/8	65.57	72.04	69.78	75.03
VT-PTQ†	4MP/8MP	65.65	72.18	69.91	75.09
LampQ	4MP/8MP	**66.27**	**72.71**	**70.24**	**75.49**
PSAQ-ViT	8/8	71.56	75.97	73.54	76.68
VT-PTQ†	8MP/8MP	71.58	76.02	73.63	76.72
LampQ	8MP/8MP	**71.77**	**76.20**	**73.76**	**76.85**

†: PSAQ-ViT quantization with bit allocation by VT-PTQ.

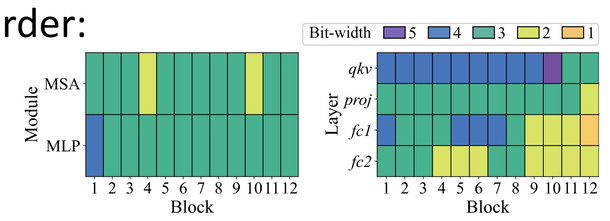
2. All ideas contribute to the improved performance

- Type-aware Fisher-based metric (I2) shows the highest impact

Method	MPQ	I1	I2	I3	Accuracy
Base: AdaLog (2024)	✗	✗	✗	✗	24.47
Base + VT-PTQ (2021b)	✓	✗	✗	✗	26.37
Base + I1 + I2	✓	✓	✓	✗	44.43
Base + I1 + I3	✓	✓	✗	✓	27.87
LampQ (Proposed)	✓	✓	✓	✓	**45.38**

3. The bit allocation reflects each layer's sensitivity

- Components closer to the input images are important
- Sensitivity by layer type in order: $qkv > fc1 \approx proj > fc2$



(a) VT-PTQ (b) LampQ (Proposed)