AdaLog: Post-Training Quantization for Vision Transformers with Adaptive Logarithm Quantizer –Supplementary Material–

Zhuguanyu Wu[1,2], Jiaxin Chen[1,2](✉), Hanwen Zhong[1,2], Di Huang[2], and Yunhong Wang[1,2]

[1] State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
[2] School of Computer Science and Engineering, Beihang University, Beijing, China
{goatwu,jiaxinchen,hanwenzhong,dhuang,yhwang}@buaa.edu.cn

In this supplementary material, we provide additional experimental results on more computer vision tasks, including object detection and instance segmentation on the COCO [3] dataset in Sec. A. Besides, we present more ablation study results of the proposed AdaLog quantizer in Sec. B.

A Experimental Results on COCO

We further evaluate our method on the object detection and instance segmentation tasks on the COCO dataset, by comparing to PTQ4VIT [5], APQ-ViT [1] and RepQ-ViT [2]. In order to make fair comparisons, we follow the experimental settings as depicted in [2], and report the AP^{box} and AP^{mask} metrics by using the Mask R-CNN and Cascade Mask R-CNN frameworks based on the Swin-T/S backbones, respectively. As shown in Table A, AdaLog consistently achieves the highest AP^{box} for object detection and AP^{mask} for instance segmentation, when performing the 6-bit quantization. Compared to the full-precision model, AdaLog incurs less than 0.6% loss in accuracy across different frameworks. For 4-bit quantization, AdaLog promotes AP^{box} by 3.0%, compared to existing approaches, when quantizing Mask R-CNN with the Swin-T backbone. Similar improvements are achieved in most cases when using the Cascade Mask R-CNN framework.

B More Ablation Study Results

B.1 Post-GELU Quantizers

In order to validate the effectiveness of the proposed AdaLog quantizer for the Post-GELU quantization, we compare with the alternative quantizers including Uniform [2], Twin Uniform [5], Log2 Quantizer [4] and $\text{Log}\sqrt{2}$ Quantizer [2]. As displayed in Table B, the compared approaches exhibit fluctuating performance for distinct architectures. For instance, the $\text{Log}\sqrt{2}$ quantizer reaches the second

✉ Corresponding author.

Table A: Comparison results on COCO for the object detection and instance segmentation tasks. AP^b and AP^m indicate AP^{box} and AP^{mask}, respectively. The best results are highlighted in bold.

Method	bits(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
PTQ4ViT [5]	4/4	6.9	7.0	26.7	26.6	14.7	13.5	0.5	0.5
APQ-ViT [1]	4/4	23.7	22.6	**44.7**	40.1	27.2	24.4	47.7	41.1
RepQ-ViT [2]	4/4	36.1	36.0	44.2	40.2	47.0	41.1	49.3	43.1
AdaLog (Ours)	4/4	**39.1**	**37.7**	44.3	**41.2**	**48.2**	**42.3**	**50.6**	**44.0**
PTQ4ViT [5]	6/6	5.8	6.8	6.5	6.6	14.7	13.6	12.5	10.8
APQ-ViT [1]	6/6	**45.4**	41.2	47.9	42.9	48.6	42.5	50.5	43.9
RepQ-ViT [2]	6/6	45.1	41.2	47.8	43.0	50.0	43.5	51.4	44.6
AdaLog (Ours)	6/6	**45.4**	**41.3**	**48.0**	**43.2**	**50.1**	**43.6**	**51.7**	**44.8**

Table B: Ablation results (%) on the post-GELU quantizers on ImageNet with the W4/A4 setting. "T-Uniform" is the abbreviation for the Twin-Uniform Quantizer in PTQ4ViT [5]. "Rep." is the abbreviation for the Bias Reparametrization. The best results are highlighted in bold.

Method	Rep.	ViT-S	ViT-B	DeiT-T	DeiT-S	DeiT-B	Swin-S	Swin-B
Full-Precision	-	81.39	84.54	72.21	79.85	81.80	83.23	85.27
Uniform [2]	×	63.14	78.08	59.93	69.23	76.02	78.79	80.67
T-Uniform [5]	×	65.29	<u>78.76</u>	60.96	69.78	76.69	<u>80.51</u>	<u>80.93</u>
Log2 [4]	✓	39.83	71.27	59.33	66.30	68.53	80.36	78.95
Log$\sqrt{2}$ [2]	✓	<u>72.44</u>	46.16	<u>62.91</u>	<u>70.60</u>	<u>77.15</u>	75.91	24.50
AdaLog	✓	**72.75**	**79.68**	**63.52**	**72.06**	**78.03**	**80.77**	**82.47**

best results with the ViT-S, Deit-T, Deit-S, and DeiT-B backbones, but degrades when quantizing ViT-B, Swin-S and Swin-B. In contrast, AdaLog steadily reaches the highest top-1 accuracy.

B.2 Post-Softmax Quantizers

Similarly, we evaluate AdaLog on post-Softmax quantization in Table C, comparing to the Log2 and Log$\sqrt{2}$ quantizers. We fix the bit-width of all other quantizers to W4A4, and compare the performance of different post-Softmax quantizers under various quantization bit-widths. Under the 4-bit setting, AdaLog achieves the best results in most cases, which are comparable to the full-precision ones. Under the 3-bit and 2-bit settings, the Log2 and Log$\sqrt{2}$ quantizers are prone to collapse with extremely low accuracy. In contrast, AdaLog performs much more steadily.

Table C: Ablation results (%) of the post-Softmax quantizers with different bit-width on ImageNet using the W4/A4 setting.

Model	W4/A4/S32	Method	W4/A4/S4	W4/A4/S3	W4/A4/S2
ViT-S/224	72.87	log2	56.74	51.88	0.10
		$\log\sqrt{2}$	54.78	0.10	0.10
		AdaLog	**72.75**	**72.39**	**70.36**
ViT-B/224	80.13	log2	78.61	76.44	0.10
		$\log\sqrt{2}$	78.91	0.10	0.10
		AdaLog	**79.68**	**79.60**	**78.38**
DeiT-T/224	63.84	log2	62.91	60.79	0.10
		$\log\sqrt{2}$	62.46	0.10	0.10
		AdaLog	**63.52**	**62.86**	**59.92**
DeiT-S/224	72.18	log2	71.83	70.91	0.10
		$\log\sqrt{2}$	71.64	0.10	0.10
		AdaLog	**72.06**	**71.35**	**69.39**
DeiT-B/224	78.29	log2	77.82	77.03	0.10
		$\log\sqrt{2}$	77.93	0.10	0.10
		AdaLog	**78.03**	**77.86**	**76.50**
Swin-S/224	81.01	log2	80.81	80.62	0.10
		$\log\sqrt{2}$	**80.77**	30.46	0.10
		AdaLog	**80.77**	**80.83**	**80.62**
Swin-B/224	82.55	log2	81.87	81.56	0.10
		$\log\sqrt{2}$	81.97	44.41	0.10
		AdaLog	**82.47**	**82.08**	**81.63**

References

1. Ding, Y., Qin, H., Yan, Q., Chai, Z., Liu, J., Wei, X., Liu, X.: Towards accurate post-training quantization for vision transformer. In: ACM MM. pp. 5380–5388 (2022)
2. Li, Z., Xiao, J., Yang, L., Gu, Q.: Repq-vit: Scale reparameterization for post-training quantization of vision transformers. In: ICCV. pp. 17227–17236 (2023)
3. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: common objects in context. In: ECCV. pp. 740–755 (2014)
4. Lin, Y., Zhang, T., Sun, P., Li, Z., Zhou, S.: Fq-vit: Post-training quantization for fully quantized vision transformer. In: IJCAI. pp. 1173–1179 (2022)
5. Yuan, Z., Xue, C., Chen, Y., Wu, Q., Sun, G.: Ptq4vit: Post-training quantization for vision transformers with twin uniform quantization. In: ECCV. pp. 191–207 (2022)