Zheyang LI, Chaoxiang LAN, Kai ZHANG, Wenming TAN, Ye REN, Jun XIAO, 2025. An adaptive outlier correction quantization method for vision Transformers. *Frontiers of Information Technology & Electronic Engineering*, 26(10):1879-1895. https://doi.org/10.1631/FITEE.2400994

An adaptive outlier correction quantization method for vision Transformers

Key words: Transformer; Model compression and acceleration; Post-training quantization; Outlier

Corresponding author: Jun XIAO
E-mail: junx@cs.zju.edu.cn
ORCID: https://orcid.org/0000-0003-0303-134X

Motivation

Transformers have demonstrated considerable success across various domains but are constrained by their significant computational and memory requirements. This poses challenges for deployment on resource-constrained devices. Quantization, as an effective model compression method, can significantly reduce the operation time of Transformers on edge devices. Notably, Transformers display more substantial outliers than convolutional neural networks, leading to uneven feature distribution among different channels and tokens. To address this issue, we propose an adaptive outlier correction quantization (AOCQ) method for Transformers.

Main idea

- We revisit the fully quantized Transformers and highlight that the key problem is the huge discrepancy in channels and tokens of the Transformers, which leads to significant accuracy degradation.

- We propose AOCQ, a simple but effective post-training quantization algorithm, which reduces the imbalance of channels and tokens at three levels: operator level, framework level, and loss level. The theoretical analysis proves that the quantization error can be reduced obviously.

Method

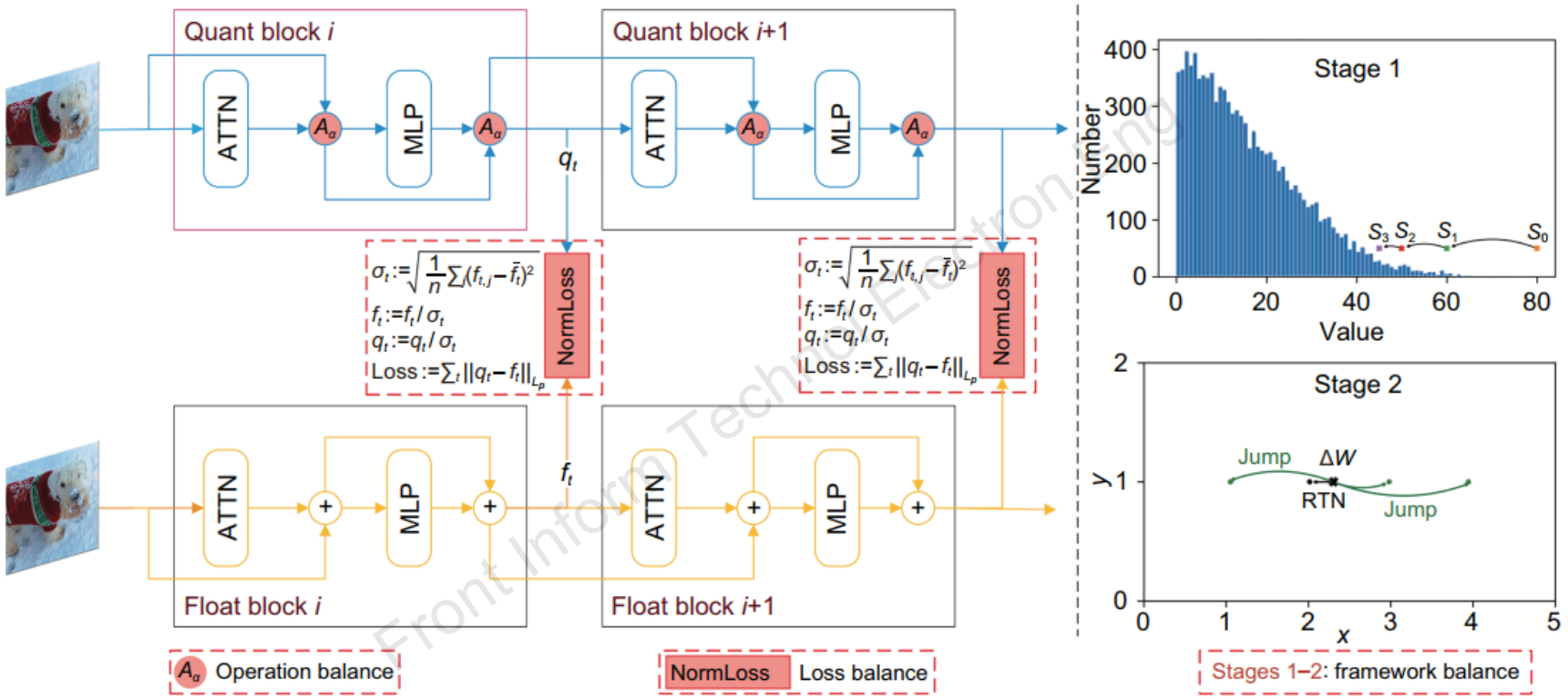


Fig. 1 Framework of the AOCQ. AOCQ deals with the outliers adaptively in three levels: operator level, framework level, and loss level. It uses Add_α and LN_α to diminish the imbalance between channels. A two-stage optimization method has been used to optimize quantization scale and the rounding direction. In the first stage, the quantization scale s_0 of activation is iteratively refined to s_4 through successive optimization steps. After that, the rounding direction of the weights is optimized in the second stage. The token norm function suppresses the extreme tokens in the block-wise reconstruction loss

Results

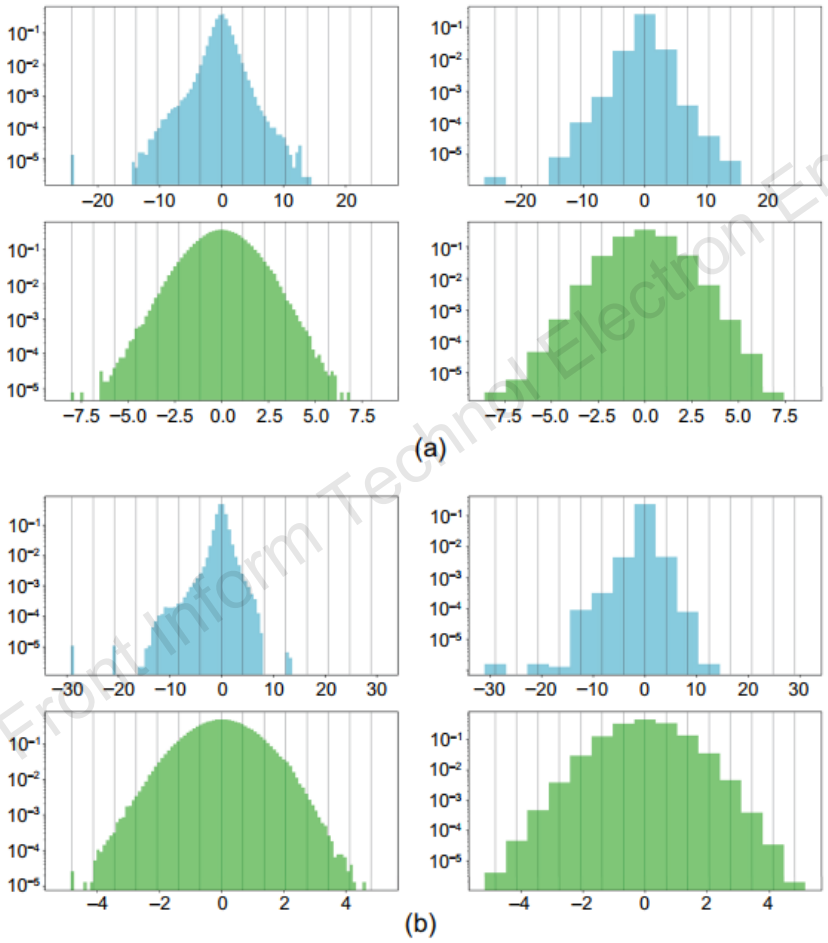


Fig. 5 Comparison of the density distribution: (a) BasePTQ; (b) our method. The x-axis represents the value interval, while the y-axis represents the count of activations falling within each interval. The left and right columns represent the original and quantized distributions, respectively

Table 2 Comparison of the performance of the proposed PTQ method AOCQ with other SOTA methods for image classification on ImageNet

Model	Method	W/A	Model size (M)	Accuracy (%)	δ	W/A	Model size (M)	Accuracy (%)	δ
Swin-Tiny		32/32	112	81.35		32/32	112	81.35	
	BasePTQ*	8/8	28	80.96	−0.39	4/8	14	66.41	−14.94
	PTQ4ViT*	8/8	28	81.24	−0.11	4/8	14	77.96	−3.39
	FQ-ViT	8/8	28	80.51	−0.84	4/8	14	78.23	−3.12
	AOCQ (ours)	8/8	28	**81.25**	**−0.10**	4/8	14	**80.87**	**−0.48**
Swin-Small		32/32	195	83.20		32/32	195	83.20	
	BasePTQ*	8/8	49	82.75	−0.45	4/8	24	78.43	−4.77
	PTQ4ViT*	8/8	49	**83.10**	**−0.10**	4/8	24	80.25	−2.95
	FQ-ViT	8/8	49	82.71	−0.49	4/8	24	81.62	−1.58
	AOCQ (ours)	8/8	49	83.00	−0.20	4/8	24	**82.80**	**−0.40**
Swin-Base		32/32	345	83.60		32/32	345	83.60	
	BasePTQ*	8/8	86	83.32	−0.28	4/8	43	78.43	−5.17
	PTQ4ViT*	8/8	86	**83.56**	**−0.06**	4/8	43	80.25	−3.35
	FQ-ViT	8/8	86	82.97	−0.63	4/8	43	81.62	−1.98
	AOCQ (ours)	8/8	86	83.50	−0.10	4/8	43	**83.42**	**−0.18**
DeiT-Tiny		32/32	22	72.21		32/32	22	72.21	
	BasePTQ*	8/8	5.6	71.28	−0.93	4/8	2.8	56.58	−15.63
	PTQ4ViT*	8/8	5.6	71.57	−0.64	4/8	2.8	66.70	−5.51
	FQ-ViT	8/8	5.6	71.61	−0.60	4/8	2.8	65.78	−6.43
	AOCQ (ours)	8/8	5.6	**71.74**	**−0.47**	4/8	2.8	**70.18**	**−2.03**
DeiT-Small		32/32	86	79.85		32/32	86	79.85	
	BasePTQ*	8/8	22	77.65	−2.20	4/8	11	64.62	−15.23
	PTQ4ViT*	8/8	22	**79.47**	**−0.38**	4/8	11	77.03	−2.82
	FQ-ViT	8/8	22	79.17	−0.68	4/8	11	75.65	−4.20
	AOCQ (ours)	8/8	22	79.19	−0.66	4/8	11	**78.60**	**−1.25**
DeiT-Base		32/32	338	81.85		32/32	338	81.85	
	BasePTQ*	8/8	84	80.94	−0.91	4/8	42	74.27	−7.58
	PTQ4ViT*	8/8	84	81.48	−0.37	4/8	42	79.59	−2.26
	FQ-ViT	8/8	84	81.20	−0.65	4/8	42	79.36	−2.49
	AOCQ(ours)	8/8	84	**81.57**	**−0.28**	4/8	42	**81.48**	**−0.37**

The input resolution is set to 224 × 224. W/A means weight/activation. * means that the quantization is only used for matrix multiplication, where add and other operations maintain float. M stands for million. BasePTQ is from Yuan ZH et al. (2022). The best results are in bold, and the second-best results are underlined

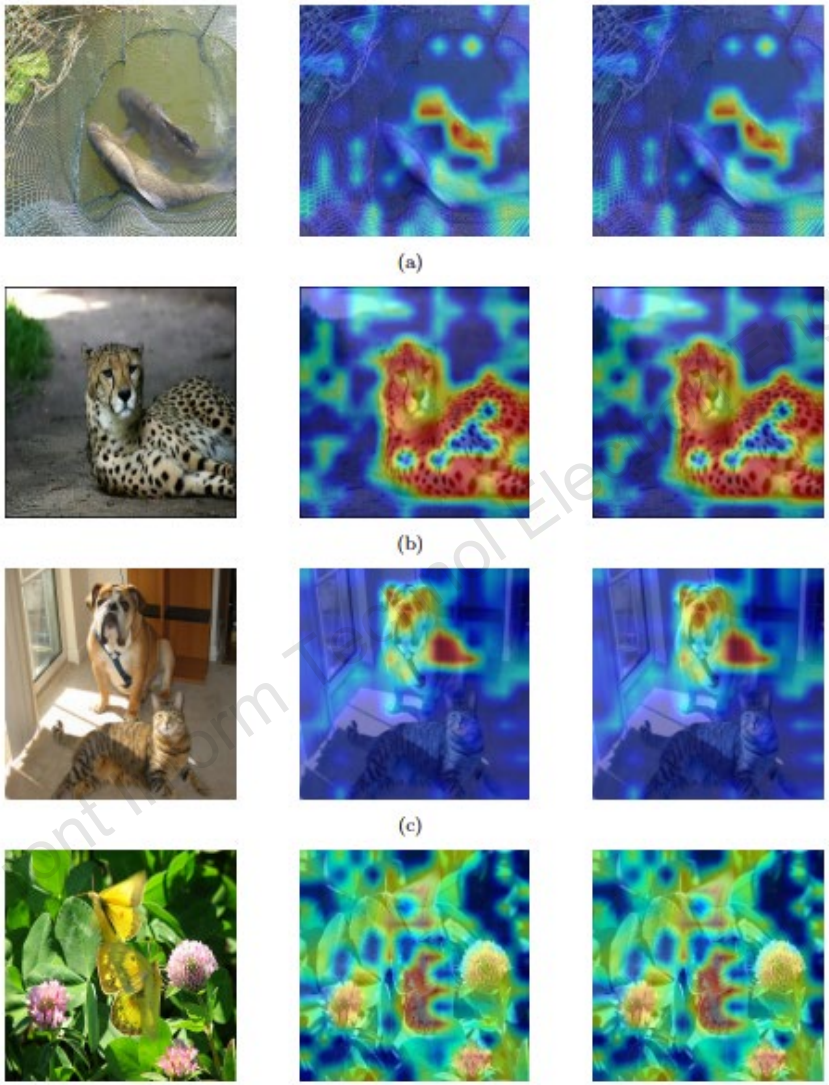


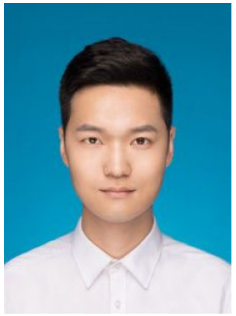
Fig. 9 Gradient-weighted class activation maps: (a) fish; (b) leopard; (c) dog; (d) butterfly. We use Grad-CAM to visualize the attention maps of the ViT-Base model. The left is the original input image. The middle is the Grad-CAM on FP32 model and the right is the Grad-CAM on W4A8 quantized model with our method. All images are resized to 224 × 224 resolutions, and the output of the last layer is used for visualization

Conclusions

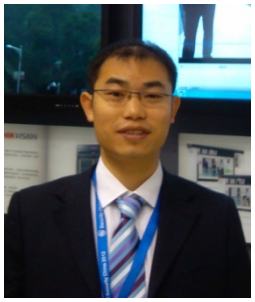
In this paper, we analyze the issue of outliers in Transformers and theoretically demonstrate that this phenomenon is caused by the structural design of the Transformer itself. To address this problem, a new method (i.e., AOCQ) has been proposed, which balances the impact of outliers from the operator level, framework level, and loss level. Based on the analysis of the condition number of the Fisher information matrix, we theoretically prove the effectiveness of our method. Experimentally, various tasks, including classification and object detection, have been validated.

Mr. Zheyang LI is an algorithm researcher at Hikvision Research Institute. He received the MS degree in Shanghai Jiao Tong University, Shanghai, China, in 2015. His current research interests include perception algorithm, neural network acceleration, and explainable AI.

Chaoxiang LAN received the BS and MS degrees from Zhejiang University, Hangzhou, China, in 2015 and 2020, respectively. He has been working in Hikvision Research Institute since September 2020. His research interests include computer vision and machine learning.



Kai ZHANG received the BS and MS degrees from the University of Science and Technology of China (USTC), China, in 2015 and 2018, respectively. He is currently with the Hikvision Research Institute, Hangzhou. His research interests include computer vision and machine learning.



Wenming TAN received his BS degree from the School of Computer Science and Technology at Dalian University of Technology in 2006 and his MS degree in circuits and systems at the University of Science and Technology of China in 2009. He has been working in Computer Vision and Machine Learning R&D at Hikvision Research Institute since 2009. His current research interests include perception algorithm and neural network acceleration.

Ye REN received her MS degree from Zhejiang University, Hangzhou, China, in 2007. She has been working at Hikvision Research Institute since 2007. Her current research interests include AI, machine perception, and robotics.

Jun XIAO received the PhD degree in computer science and technology from the College of Computer Science and Technology, Zhejiang University, Hangzhou, China, in 2007. He is currently a professor with Zhejiang University. His research interests include computer vision, multimedia retrieval, and machine learning.