Outlier-Aware Slicing for Post-Training Quantization in Vision Transformer

Yuexiao Ma [1 2] **Huixia Li** [3] **Xiawu Zheng** [2 4 5] **Feng Ling** [3] **Xuefeng Xiao** [3] **Rui Wang** [3] **Shilei Wen** [3] **Fei Chao** [2] **Rongrong Ji** [2 5]

Abstract

Post-Training Quantization (PTQ) is a vital technique for network compression and acceleration, gaining prominence as model sizes increase. This paper addresses a critical challenge in PTQ: **the severe impact of outliers on the accuracy of quantized transformer architectures.** Specifically, we introduce the concept of ‘reconstruction granularity’ as a novel solution to this issue, which has been overlooked in previous works. Our work provides theoretical insights into the role of reconstruction granularity in mitigating the outlier problem in transformer models. This theoretical framework is supported by empirical analysis, demonstrating that varying reconstruction granularities significantly influence quantization performance. Our findings indicate that different architectural designs necessitate distinct optimal reconstruction granularities. For instance, the multi-stage Swin Transformer architecture benefits from finer granularity, a deviation from the trends observed in ViT and DeiT models. We further develop an algorithm for determining the optimal reconstruction granularity for various ViT models, achieving state-of-the-art (SOTA) performance in PTQ. For example, applying our method to 4-bit quantization, the Swin-Base model achieves a Top-1 accuracy of 82.24% on the ImageNet classification task. This result surpasses the RepQ-ViT by 3.92% (82.24% VS 78.32%). Similarly, our approach elevates the ViT-Small to a Top-1 accuracy of 80.50%, outperforming NoisyQuant by 3.64% (80.50% VS 76.86%).

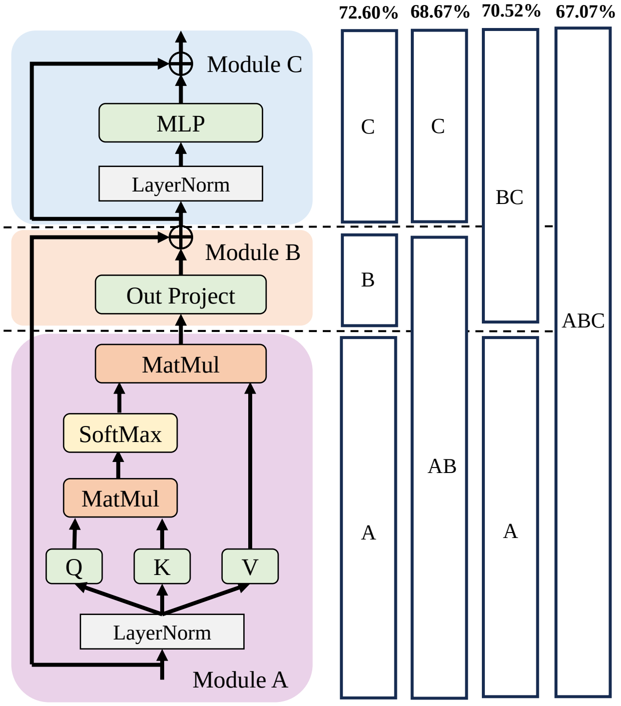


Figure 1. Introducing finer-grained slicing involves dividing the transformer block into three modules: self-attention module A, out project module B, and MLP module C. Different combination granularity leads to significant differences in accuracy.

1. Introduction

With the development of neural networks, transformer-like structures (Dosovitskiy et al., 2020; Touvron et al., 2021; Liu et al., 2021a) dominate on various vision tasks. As transformer blocks are stacked and model dimensions increase, the performance of transformer-like models continues to improve (Dosovitskiy et al., 2020). However, larger models require more computational resources despite their benefits. Consequently, researchers are increasingly focusing on compression methods for transformer models to reduce the model size and speed up inference, such as neural architecture search (Li et al., 2021a; Chen et al., 2021c;b;a), pruning (Dong et al., 2023; Yang et al., 2021; Tang et al., 2022; Yang et al., 2023; Yu & Xiang, 2023), and model

[1] This work was done when Yuexiao Ma was intern at ByteDance Inc. [2] Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, School of Informatics, Xiamen University, 361005, P.R. China. [3] ByteDance Inc. [4] Peng Cheng Laboratory, Shenzhen, China. [5] Institute of Artificial Intelligence, Xiamen University. Correspondence to: Rongrong Ji <rrji@xmu.edu.cn>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

quantization (Lin et al., 2021; Yuan et al., 2022; Ding et al., 2022; Li et al., 2023; Liu et al., 2023b; Sun et al., 2022).

Quantization is a technique that converts model weights and activations from floating-point numbers to low-bit fixed-point numbers (Krishnamoorthi, 2018), resulting in smaller model sizes and improved access efficiency. When implemented on hardware that supports low-bit arithmetic, quantization models can further accelerate inference. **Q**uantization-**A**ware **T**raining (QAT) (Choi et al., 2018; Gong et al., 2019; Esser et al., 2019; Bondarenko et al., 2023) and **P**ost-**T**raining **Q**uantization (PTQ) (Li et al., 2021b; Nagel et al., 2020; Wei et al., 2022a; Hubara et al., 2021; Lin et al., 2021; Yuan et al., 2022; Ding et al., 2022; Li et al., 2023; Liu et al., 2023b) are two classes of quantization algorithms, differing in the overhead required to obtain a quantization model. PTQ is particularly favored for large-scale models due to its low data dependency and minimal algorithmic overhead. However, the efficiency of PTQ comes at the expense of a significant loss in model performance. This loss is especially pronounced in vision transformers due to the challenge posed by outliers in the data distribution (Bondarenko et al., 2023; Lin et al., 2021; Yuan et al., 2022; Ding et al., 2022; Li et al., 2023; Liu et al., 2023b), which adversely affects PTQ, particularly uniform quantization. Recent research in PTQ for vision transformers aims to address this issue. FQ-ViT (Lin et al., 2021) addresses attention map distribution pathology by employing the log2 quantizer and proposes an integer approximation of the exponential function for softmax quantization. Additionally, by combining it with power-of-two quantization of the layernorm layer (Ba et al., 2016), FQ-ViT achieves integer-only inference for vision transformers. PTQ4ViT (Yuan et al., 2022) introduces twin uniform quantization for activation values after softmax and activation functions to handle outlier distribution. APQ-ViT (Ding et al., 2022) utilizes a bottom elimination function to prioritize the loss generated by outliers during calibration. RepQ-ViT (Li et al., 2023) proposes $\log\sqrt{2}$ quantization to better adapt to the distribution of softmax output activations and employs a reparameterization technique to transition from $\log\sqrt{2}$ quantization and the per-channel activation quantization to log2 quantization and per-layer activation quantization after layernorm, resulting in improved inference efficiency. Quantizable Transformers (Bondarenko et al., 2023) employs clipped softmax to prevent gradient accumulation in pre-training and introduces gated attention to selectively update the representation of specific tokens, making the pre-trained model outlier-free and more suitable for quantization.

However, the importance of reconstruction granularity in vision transformer quantization has been overlooked by existing approaches. MRECG (Ma et al., 2023) discusses the role of reconstruction granularity in mitigating loss oscillations in convolutional neural networks. However, this assumption of topological homogeneity does not hold for different modules within transformer blocks and across transformer blocks. Therefore, it is necessary to investigate the properties of vision transformers and explore their impact on quantization performance.

In this study, we present a theorem demonstrating that coarser reconstruction granularity leads to smaller quantization loss for the same input. We conduct experiments by sampling various reconstruction granularities in different vision transformer quantization models. The experimental results confirm our theorem, as the number of transformer block combinations exhibits a negative correlation with the final block loss. However, the downsampled blocks in Swin transformer result in outliers appearing exclusively in the third stage. Consequently, the joint optimization of modules across stages only yields sub-optimal results. As a result, the quantization loss of the last transformer block and the final loss in Swin transformers show a negative correlation due to inadequate optimization, which is contrary to ViT and DeiT conclusions. Consequently, we propose two rules to establish the reconstruction granularity for various vision transformers. Furthermore, we partition the transformer block internally into three modules: self-attention, out projection, and MLP. We exhaustively compare the performance of different combinations of modules due to violation of the topological homogeneity assumption. Finally, we validate the effect of reconstruction granularity on quantization performance across different vision transformer models.

In summary, our contribution is as follows.

- We first examine the impact of outliers on quantization performance in vision transformers. (Section 3.1) And then, we propose a novel optimization approach, termed ‘Reconstruction Granularity’. Our theoretical and empirical analyses reveal the relationship between reconstruction granularity and outliers. As a result, reconstruction granularity substantially affects quantization performance. (Section 3.2)

- We offer two rules for setting the ideal granularity in diverse vision transformer models. With an assumption of transformer-like topological homogeneity, coarser granularity leads to lower quantization loss. (Section 3.3)

- We empirically validate the impact of reconstruction granularity on quantization performance across various models using the ImageNet dataset. Notably, with a $4/4$ bit quantization on DeiT-tiny, we attain a Top-1 accuracy of 66.31%. Furthermore, our approach achieves a Top-1 accuracy of 80.50% on ViT-small, surpassing NoisyQuant by a margin of 3.64% (80.50% versus 76.86%). (Section 4)

2. Related Work

Post-training quantization. PTQ (Frumkin et al., 2023; Xu et al., 2023; Lin et al., 2023a; Bai et al., 2022; Liu et al., 2023a; Jeon et al., 2022; Li & Gu, 2023) has less data dependency and higher algorithmic efficiency than QAT (Le & Li, 2023; Wang et al., 2022b; He et al., 2023; Wang et al., 2022a; Li et al., 2022a). Therefore, PTQ is preferred for rapid deployment scenarios of quantization models and large-scale model quantization scenarios. In the field of convolutional neural networks, Adaround (Nagel et al., 2020) expands the optimization space of weight rounding by an adaptive rounding technique. Further, the quantization loss at each layer is used to guide the optimization of the weight rounding parameters. BRECQ (Li et al., 2021b) further expands the layer-by-layer optimization granularity to block-by-block optimization and uses diagonal Fischer matrices to approximate the Hessian matrices for the efficiency-performance trade-off. Qdrop (Wei et al., 2022a) randomly discards quantization activations during optimization to achieve better performance. PTQ in vision transformers targets two main issues: the heavy-tailed activation distribution post-softmax and outliers in deep model blocks. Research focuses on solutions involving pre-trained models, quantization functions, and algorithms. For more details, refer to Section 1.

Granularity Reconstruction. MRECG (Ma et al., 2023) has investigated the importance of various joint optimization schemes in addressing loss oscillations in convolutional neural networks. However, the inclusion of self-attention in vision transformer models significantly distinguishes their structure from convolutional neural networks. Furthermore, vision transformers encounter the issue of outliers, which is not typically observed in convolutional neural networks. Consequently, it is crucial to examine the impact of reconstruction granularity on vision transformer models to tackle the outlier problem effectively. NWQ (wang et al., 2022) similarly uses a coarser reconstruction granularity for convolutional neural networks and proposes Activation Regularity, ASoftmax, and AMixup to solve overfitting and discrete optimization problems.

3. Methodology

In this section, we first discuss the impact of outliers in the vision transformer on quantization in Section 3.1. Secondly, we investigate the relationship between model outliers and quantization loss. Through theoretical analysis and sampling experiments, we demonstrate the importance of reconstruction granularity in optimizing outliers in Section 3.2. At last, we uncover the issue of inadequate optimization due to stages acrossing without satisfying the topological homogeneity assumption in Section 3.3, which inspires us to propose two rules for determining different reconstruction granularity for various models.

3.1. Outliers in Vision Transformer

In this subsection, we reveal the outlier problem in vision transformers and analyze how outliers affect quantization. Quantization converts floating-point numbers to low-bit fixed-point numbers, enhancing access efficiency and reducing model size. However, this efficiency improvement is accompanied by a loss in quantization model performance. The quantization function is defined as follows,

$$Q(x) = \left(clip \left(\left\lfloor \frac{x}{s} \right\rceil + zp, 0, 2^n - 1 \right) - zp \right) * s. \quad (1)$$

Where $clip(\cdot)$ denotes the truncation operation. s, zp, and n stand for the step size, zero points, and bit-width, respectively. $\lfloor \cdot \rceil$ denotes the rounding operation. To handle the non-differentiability of the rounding operation, we employ the **S**traight-**T**hrough **E**stimator (STE) (Bengio et al., 2013) for obtaining an approximate back-propagated gradient. Equation (1) shows that the quantization error comprises the truncation error beyond the domain and the rounding error within the domain. Consequently, a theoretical minimum quantization error exists when the input x follows the uniform distribution.

However, outliers in activation values are observed in the vision transformer models. Figure 4 illustrates the distribution of hidden channel data for a specific token of output activation in the 6-th and 5-th blocks of the ViT and Swin models, respectively. These distributions reveal that certain channels exhibit magnitudes significantly larger than those of other channels. Moreover, these magnitudes tend to increase in subsequent transformer blocks. Since the distribution of quantization fixed points is uniform, when outliers are not truncated by the largest quantization fixed point, the activation of the remaining channels gets massively mapped to the zero point, resulting in substantial rounding errors. In scenarios where the distribution of quantization fixed points does not account for the presence of outliers, the quantization error is primarily dominated by truncation error. Consequently, only a small fraction of activations are mapped to quantization fixed points that are uniformly distributed between outliers and non-outliers. The existence of outliers "compresses" the representation of the remaining values, posing a challenge for quantization.

3.2. The Relationship between Outliers and Reconstruction Granularity

We examine the relationship between reconstruction granularity and outliers. Initially, Figure 3 shows how outliers amplify quantization loss. Subsequently, we prove, both theoretically and through extensive experiments, that adopting coarser granularity under specific conditions can diminish quantization loss during optimization. Thus, coarser granularity helps mitigate the impact of outliers.

Outliers pose a significant challenge in post-training quantization for vision transformers. However, the reconstruction methods used in post-training quantization for convolutional neural networks may not be directly suitable for vision transformers. These methods typically optimize an intermediate agent guided by an objective function to minimize the quantization error. In the context of vision transformers, if we consider transformer blocks as the reconstruction granularity, the quantization error can be defined as follows,

$$\mathcal{L}(X_{i-1}, W_i) = \mathbb{E}\left[\|f_i(X_{i-1}, W_i) - f_i(\hat{X}_{i-1}, \hat{W}_i)\|_F^2\right]. \tag{2}$$

Where $f_i(\cdot)$ is the i-th transformer block function. X_{i-1}, W_i are the input activations and the weights of the i-th transformer block, respectively. $\hat{X}_{i-1}, \hat{W}_i$ are the corresponding quantized versions. $||\cdot||_F$ is the frobenius norm. Notably, the quantization loss demonstrates a significant decrease when outliers are removed. We evaluate the mean square error loss of the 11-th transformer block output activation of ViT-small before and after quantization according to Equation (2). To determine the activation values used for calculating the quantization error loss, we establish 20 quantile points based on the distribution interval of the absolute difference pre- and post-quantization. Any activation values exceeding a quantile point are not considered in the loss calculation. Figure 3 illustrates that excluding 5% of outliers results in a sharp two-order-of-magnitude reduction in quantization error loss. Furthermore, as more values are excluded, the loss decreases exponentially. Consequently, optimizing quantization loss is crucial for addressing the outlier problem in vision transformers.

MRECG (Ma et al., 2023) discusses the quantization loss of convolutional neural networks as affected by module capacity under the assumption of module topology homogeneity. The incorporation of self-attention in vision transformers alters the model structure, deviating from a stack of convolutional layers. Therefore, we propose the concept of "Transformer-Like Topological Homogeneity" to classify a specific group of transformer blocks.

Definition 3.1. ***(Transformer-Like Topological Homogeneity)*** Suppose two blocks have the same number of post-fusion linear layers and the operators between the corresponding linear layers of the two blocks remain consistent. Then we claim that two modules containing at least one such block are topologically homogeneous.

In Definition 3.1, we emphasize the post-fusion since successive linear layers in parallel or series can be fused into a single linear layer. Transformer blocks comprising self-attention and MLP modules exhibit topological homogeneity. This homogeneity persists whether a stack of n or m transformer blocks forms separate modules. Subsequently, we introduce a theorem that delineates the connection between reconstruction granularity and quantization loss.

Theorem 3.2. *Assuming that the two modules satisfy the transformer-like topological homogeneity, for the same input activation, the coarser granularity module corresponds to a smaller quantization error. Formally,*

$$\begin{aligned}&\mathcal{L}(X_i, W_{i+1}, \cdots, W_{i+n})\\ &\leq \mathcal{L}(X_i, W^{'}_{i+1}, \cdots, W^{'}_{i+h}, \cdots, W_{i+n}).\end{aligned} \tag{3}$$

Where $\{W^{'}_{i+1}, \cdots, W^{'}_{i+h}\}$ *make* $\mathcal{L}(X_i, W^{'}_{i+1}, \cdots, W^{'}_{i+h})$ *optimal.*

Please refer to Appendix A for comprehensive proof. Theorem 3.2 establishes that a coarser granularity results in a reduced quantization loss during optimization. Furthermore, we conduct extensive sampling of reconstruction granularities on ViT, DeiT, and Swin models. Given the complexity of the candidate space, our sampling is limited to combinations of up to three consecutive transformer blocks. However, for Swin transformers, we break the definition of transformer-like topological homogeneity, which constrains the sampling space due to the downsampling block. This implies allowing cross-stage combinations of transformer blocks in the Swin model. Random sampling is performed to obtain a substantial number of reconstruction granularity schemes, and optimization is utilized to restore quantization accuracy. As shown in the first row of Figure 2, the number of combinations is negatively correlated with the final block loss on all vision transformers. This experimentally indicates that coarser reconstruction granularity leads to smaller model quantization loss. In essence, coarser reconstruction granularity effectively addresses the outlier problem in vision transformers.

3.3. Granularity Determination

Based on our analysis, we summarize two key rules:

Rule 1: For vision transformers with transformer-like topological homogeneity, coarse granularity quantization is recommended.

Rule 2: Without such homogeneity, finer granularity quantization is favored.

In the case of the Swin transformer, the presence of outliers is observed primarily in the third stage, which contains a large stack of transformer blocks. By adhering to the definition of transformer-like topological homogeneity and limiting the search within stages, the Swin model yields consistent results with ViT and DeiT models. Please refer to Appendix B for detailed experimental information. However, the search results are confined to a local optimum with the in-stage restriction. On the other hand, a global search for the Swin model leads to inadequate optimization due to the violation of the definition of topological homogeneity. Specifically, when the reconstruction module traverses the h-th and $(h+1)$-th stages of Swin, the gradient of the parameter at stage h is defined as follows:

$$\begin{aligned}
&\frac{\partial \mathcal{L}(X_{i-1}^{h+1}, W_i^{h+1})}{\partial W_j^h} \\
&= \mathbb{E}\Big[2\Big\|f_i^{h+1}(X_{i-1}^{h+1}, W_i^{h+1}) - f_i^{h+1}(\hat{X}_{i-1}^{h+1}, \hat{W}_i^{h+1})\Big\|_F \\
&\times \Big(f_i^{'h+1}(X_{i-1}^{h+1}, W_i^{h+1}) - f_i^{'h+1}(\hat{X}_{i-1}^{h+1}, \hat{W}_i^{h+1})\Big) \\
&\times f_{DS}^{'}(X_l^h, W_{DS}) \times \prod_{a=1}^{i-1} f_a^{'h+1}(X_{a-1}^{h+1}, W_a^{h+1}) \\
&\times \prod_{b=j}^{l} f_b^{'h}(X_{b-1}^h, W_b^h) \times \frac{\partial f_j^h(X_{j-1}^h, W_j^h)}{\partial W_j^h}\Big].
\end{aligned} \tag{4}$$

Where superscript represents the stage and subscript represents the transformer block. For example, $f_b^{'h}$ denotes the b-th transformer block of the h-th stage in the Swin takes the partial derivative with respect to the activation. x_{i-1}^{h+1} refers to the output activation of the $(i-1)$-th transformer block of the $(h+1)$-th stage. The term f_{DS} represents the downsampled block. l denotes the number of transformer blocks in the h-th stage. If under the definition of transformer-like topological homogeneity, the optimization parameters in stage h are not affected by the quantization loss in stage $h+1$. Specifically, the gradient of the parameters of the module within stage h is,

$$\begin{aligned}
&\frac{\partial \mathcal{L}(X_{e-1}^{h}, W_e^{h})}{\partial W_j^h} \\
&= \mathbb{E}\Big[2\Big\|f_e^{h}(X_{e-1}^{h}, W_e^{h}) - f_e^{h}(\hat{X}_{e-1}^{h}, \hat{W}_e^{h})\Big\|_F \\
&\times \Big(f_e^{'h}(X_{e-1}^{h}, W_e^{h}) - f_e^{'h}(\hat{X}_{e-1}^{h}, \hat{W}_e^{h})\Big) \\
&\times \prod_{b=j}^{e-1} f_b^{'h}(X_{b-1}^h, W_b^h) \times \frac{\partial f_j^h(X_{j-1}^h, W_j^h)}{\partial W_j^h}\Big].
\end{aligned} \tag{5}$$

Where e is the e-th transformer block of the h-th stage and $e <= l$. From Equations (4) and (5), we note that crossing stages determines if the optimized parameter gradient at stage h is affected by the gradient at stage $h+1$ and the downsampled layer. There's also a marked change in the Frobenius norm of differences pre- and post-quantization between stages. Specifically, the norm magnitude during the outlier accumulation stage significantly exceeds that of the non-outlier stage, impacting parameter optimization at stage h negatively. The specific derivation of Equations (4) and (5), along with further experimental analysis, can be found in Appendix C. In the bottom line of Figure 2, we observe a consistent positive correlation between the quantization loss and the final performance of the ViT and DeiT. This indicates that using a coarser reconstruction granularity leads to improved quantization model performance. We present quantization loss distributions for various granularity of ViT and DeiT models in our ablation study. In addition, our search space is limited to combinations of up to 3 consecutive transformer blocks. However, it should be noted that the more coarse granularity results in compromised accuracy. For a more comprehensive analysis, please refer to Section 4.3. For the Swin model, the problem of inadequate optimization leads to opposite conclusions from ViT and DeiT. Specifically, finer granularity corresponds to better quantization performance. This observation motivates us to explore further slicing of the transformer block.

We partition the transformer block into three modules: self-attention module A, out projection module B, and MLP module C. We explore various combinations of these modules across different models. Notably, the optimization within the transformer blocks of ViT and DeiT models does not adhere to the principle that coarser granularity yields superior quantization performance, as transformer-like topological homogeneity is not satisfied. Conversely, the Swin model achieves optimal quantization performance when employing the finest granularity slice, corroborating our earlier findings. Please refer to Section 4.3 and Appendix D for detailed experiments and algorithm.

4. Experiments

In this section, we aim to validate the effectiveness of our algorithm. First, we provide comprehensive details of the experimental setup. Next, we conduct a comparative analysis between our method and a wide range of post-training quantization methods for vision transformer models on ImageNet. Finally, we perform an ablation study of our method, which includes validation at different granularities, exploration of various combinations of internal slices within transformer blocks, and comparisons of quantization loss between different granularity optimizations.

4.1. Settings

For the reconstruction optimization, we incorporate the optimization process in Adaround (Nagel et al., 2020) for weight rounding. In addition, we also refer to the random drop method for activations in QDrop (Wei et al., 2022a). We take 16 batch data for PTQ optimization, and the batch size is 64. For the hyper-parameter settings of the optimization parameters, such as reconstruction iteration, learning rate, etc., we refer to the default settings of the above methods and keep them consistent. Please refer to Appendix F for details. We conduct our experiments on NVIDIA Tesla

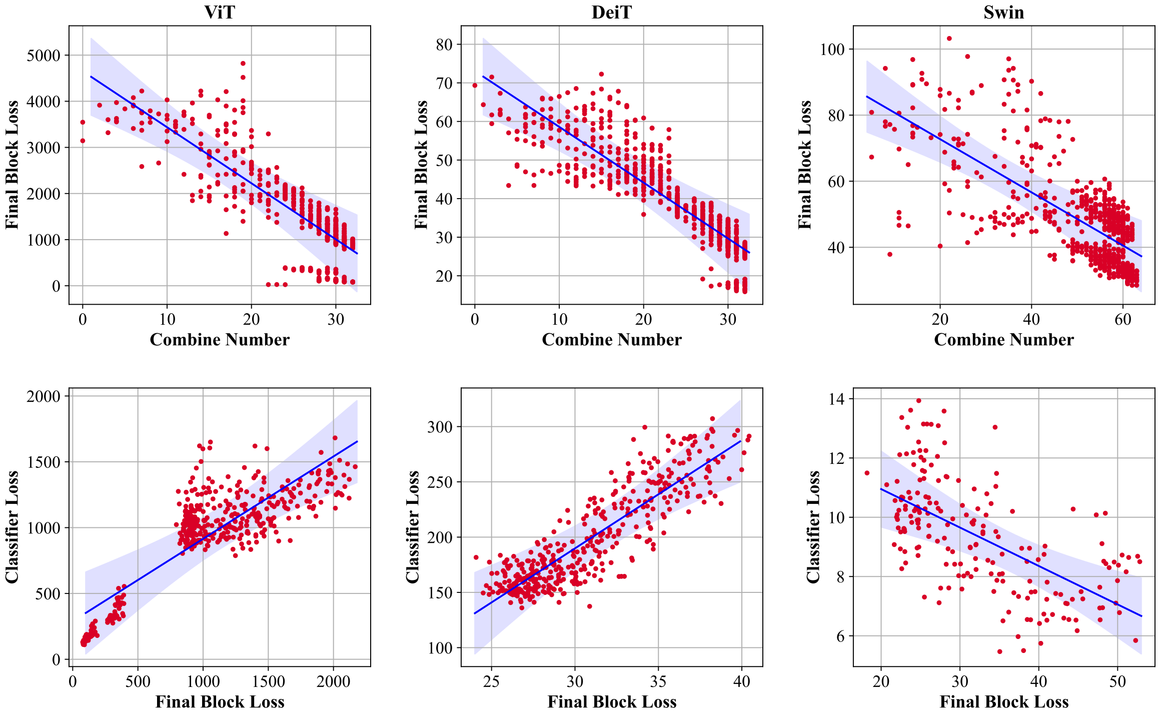


Figure 2. The relationship between reconstruction granularity and quantization model performance. The "Combine Number" indicates the number of two consecutive transformer block combinations, reflecting the coarseness of the reconstruction granularity. The "Final Block Loss" represents the output quantization loss of the last transformer block, while "Classifier Loss" denotes the final quantization loss of the model, which correlates positively with model performance. The top line illustrates the relationship between reconstruction granularity and model quantization loss, while the bottom line illustrates the relationship between model quantization loss and model performance.

Table 1. Comparison of the Top-1 accuracy (%) of our algorithm with the State-Of-The-Art method on ImageNet. "T", "S", and "B" represent the Tiny, Small, and Base models, respectively. "W/A" denotes the bit-width of the weights and activation quantization.

Methods	W/A	ViT-S	ViT-B	DeiT-T	DeiT-S	DeiT-B	Swin-S	Swin-B
FP32	–	81.38	84.53	72.13	79.83	81.80	83.23	85.27
FQ-ViT (Lin et al., 2021)	4/4	0.1	0.1	0.1	0.1	0.1	0.1	0.1
PTQ4ViT (Yuan et al., 2022)	4/4	42.57	30.69	36.96	34.08	64.39	76.09	74.02
APQ-ViT (Ding et al., 2022)	4/4	47.95	41.41	47.94	43.55	67.48	77.15	76.48
RepQ-ViT (Li et al., 2023)	4/4	65.05	68.48	57.43	69.03	75.61	79.45	78.32
Ours	4/4	72.88	76.59	66.31	76.00	78.83	81.02	82.46
FQ-ViT (Lin et al., 2021)	6/6	4.26	0.1	58.66	45.51	64.63	66.50	52.09
PSAQ-ViT (Li et al., 2022b)	6/6	37.19	41.52	57.58	63.61	67.95	72.86	76.44
Ranking (Liu et al., 2021c)	6/6	–	75.26	–	74.58	77.02	–	–
PTQ4ViT (Yuan et al., 2022)	6/6	78.63	81.65	69.68	76.28	80.25	82.38	84.01
NoisyQuant (Liu et al., 2023b)	6/6	78.65	82.32	–	77.43	80.70	82.86	84.68
APQ-ViT (Ding et al., 2022)	6/6	79.10	82.21	70.49	77.76	80.42	82.67	84.18
RepQ-ViT (Li et al., 2023)	6/6	80.43	83.62	70.76	78.90	81.27	82.79	84.57
Ours	6/6	80.60	83.81	71.52	79.50	81.72	82.76	84.91

V100.

Regarding the reconstruction granularity, we adopt a combination strategy involving three consecutive transformer blocks for ViT and DeiT models. For the Swin model, we employ the finest-grained scheme, optimizing the three modules individually as depicted in Figure 1. Furthermore, as the coarser-grained gains demonstrated in Figure 2 do not seem

Table 2. Different module combinations within a transformer block are validated using the following configurations: A-B-C, AB-C, A-BC, and ABC. In the A-B-C configuration, modules A, B, and C were not combined during the optimization process. AB-C and A-BC configurations involved joint optimization of module A with module B and module B with module C, respectively. The ABC configuration optimized each transformer block independently. All experiments quantized both the weight and activation of the model to 4 bit.

Model	FP32	A-B-C	AB-C	A-BC	ABC
ViT-Small	81.38	72.60	68.67	70.52	67.07
DeiT-Tiny	72.13	64.07	64.30	64.06	64.71
Swin-Small	83.23	80.98	80.89	80.57	79.60

Table 3. Validation of extensive reconstruction granularity. We validate a wider range of reconstruction granularity on ViT-Small and DeiT-Tiny. “n-block” implies that the model performs a combination of n consecutive transformer blocks during the optimization process.

Model	FP32	1-block	2-block	3-block	4-block	6-block	12-block
ViT-Small	81.38	67.38	72.86	73.32	71.81	70.31	66.33
DeiT-Tiny	72.13	64.89	66.15	66.31	65.91	65.11	60.15

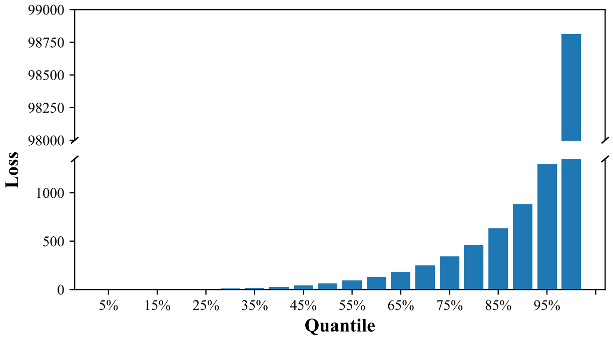


Figure 3. Outliers contribute significantly to the quantization error loss. Specifically, we analyze the mean square error loss of the 11-th transformer block output of ViT-small before and after quantization. Additionally, we vary the proportions of values retained for the loss calculation by using quantile points. Any values beyond the quantile points are assigned a value of 0.

to reach saturation for ViT and DeiT models, we conduct additional ablation experiments with the more coarse-grained settings. For further information, please refer to Section 4.3 for a detailed analysis.

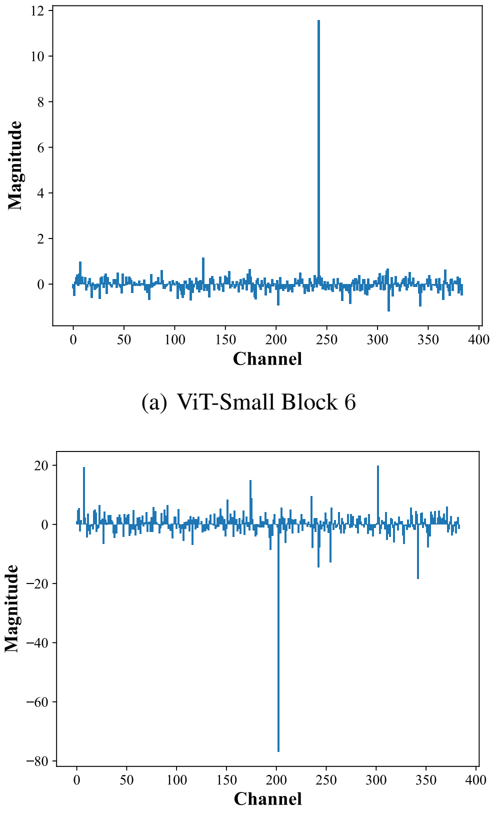
4.2. Main Results

Table 1 presents a comprehensive comparison of our algorithm with other post-training quantization methods for vision transformers across various models and quantization configurations. Notably, we outperform RepQ-ViT by 8.88% on the 4/4 bit quantization of DeiT-Tiny, achieving an accuracy of 66.31% compared to their 57.43%. Similarly, our method achieves an accuracy of 72.88% on the 4/4 bit quantization of ViT-small, surpassing RepQ-ViT by 7.83% (72.88% vs. 65.05%). Furthermore, our method demonstrates more pronounced gains in low-bit quantization. For instance, the accuracy improvement on the 4/4 bit quantization of ViT-small (7.83%) exceeds that of the 6/6 bit quantization (0.17%). Since low-bit quantization is more susceptible to outliers due to the limited number of quantized fixed points, our method effectively addresses the outlier problem through optimal granularity, resulting in significant performance enhancements.

4.3. Ablation Study

Finer reconstruction granularity. We examine the impact of combining modules with finer reconstruction granularity within a transformer block across different models in Table 2. Contrary to previous findings, the notion that coarser granularity leads to superior quantization performance in ViT and DeiT is no longer valid due to the absence of transformer-like topological homogeneity. This is evident in the performance of ViT-Small, where the finer-grained A-B-C scheme is outperformed by the coarser-grained scheme ABC (72.60% vs. 67.07%). Additionally, in the presence of a downsampled layer, a globally optimal sampling strategy is maintained for the Swin model as illustrated in Figure 2. The Swin model’s preference for finer-grained reconstruction schemes aligns with the experimental results presented in Table 2. Furthermore, as outliers tend to appear in modules B and C, the joint optimization of modules AB in ViT still suffers from inadequate optimization, resulting in lower quantization accuracy.

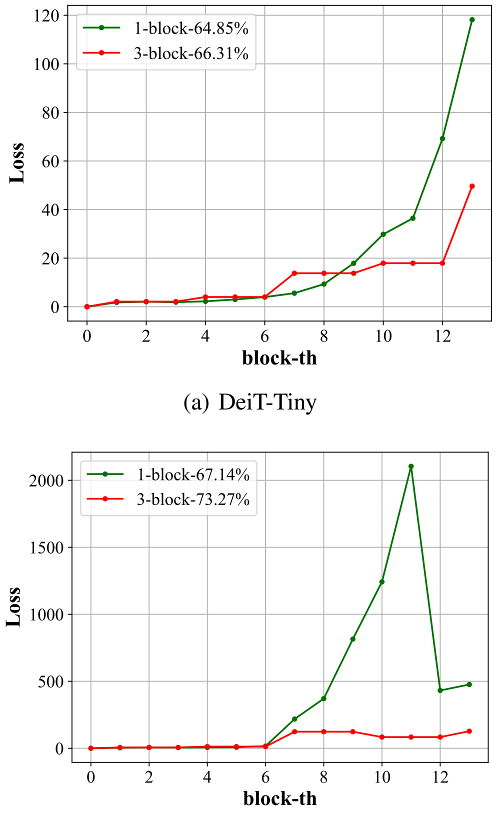
Coarser reconstruction granularity. To assess the impact of coarser reconstruction granularity on ViT and DeiT models, we conduct experiments as the quantization performance gains demonstrated in Figure 2 do not reach saturation. During the optimization process, we combine consecutive n transformer blocks (where $n \in \{1, 2, 3, 4, 6, 12\}$)



(a) ViT-Small Block 6

(b) Swin-Small Block 5

Figure 4. Outliers in vision transformers. We extract the output activations of the 6-th and 5-th transformer blocks in the ViT and Swin models, respectively. The magnitude of the hidden channel data distribution is recorded for a specific token within the activation value.



(a) DeiT-Tiny

(b) ViT-Small

Figure 5. Loss distribution and Top-1 accuracy at different granularities for ViT and DeiT. “1-block” denotes single transformer block optimization. “3-block” denotes the joint optimization of three consecutive transformer blocks.

for joint optimization. The results in Table 3 consistently indicate a decrease in quantization performance for both ViT and DeiT models. When the reconstruction granularity reaches a certain threshold, the parameter optimization resembles the process of quantization-aware training (QAT). Fine-tuning for quantization error in QAT often requires a large amount of data, whereas the limited data available for post-training quantization can lead to overfitting of the optimized parameters when the reconstruction granularity is too large. In summary, the quantization performance of ViT and DeiT models initially improves but eventually declines as the reconstruction granularity becomes coarser.

Loss distribution. We optimize ViT-Small and DeiT-Tiny using different granularities. Figure 5 presents the loss distribution and Top-1 accuracy of ViT and DeiT models at different granularities, demonstrating that optimizing with a coarser granularity significantly reduces quantization loss and addresses the outlier problem in vision transformers.

5. Conclusion

In this paper, we investigate the problem of outliers in the vision transformer model and its impact on quantization. We also establish the equivalence between optimizing the quantization loss and addressing the outlier problem through a quantile exclusion experiment. By converting the contradiction from the vision transformer model’s outlier problem to the optimization problem of quantization loss, we demonstrate, under the transformer-like topological homogeneity, that coarser-grained reconstruction effectively reduces the quantization loss of ViT and DeiT models. We extensively validate the effectiveness of our algorithm across various quantization configurations for numerous vision models.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgments

This work was supported by National Science and Technology Major Project (No. 2022ZD0118202), the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No. U21B2037, No. U22B2051, No. 62176222, No. 62176223, No. 62176226, No. 62072386, No. 62072387, No. 62072389, No. 62002305 and No. 62272401), and the Natural Science Foundation of Fujian Province of China (No.2021J01002, No.2022J06001).

References

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Bai, H., Hou, L., Shang, L., Jiang, X., King, I., and Lyu, M. R. Towards efficient post-training quantization of pre-trained language models. *Advances in Neural Information Processing Systems*, 35:1405–1418, 2022.

Bengio, Y., Léonard, N., and Courville, A. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Bondarenko, Y., Nagel, M., and Blankevoort, T. Quantizable transformers: Removing outliers by helping attention heads do nothing. *arXiv preprint arXiv:2306.12929*, 2023.

Chen, B., Li, P., Li, C., Li, B., Bai, L., Lin, C., Sun, M., Yan, J., and Ouyang, W. Glit: Neural architecture search for global and local image transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12–21, 2021a.

Chen, M., Peng, H., Fu, J., and Ling, H. Autoformer: Searching transformers for visual recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12270–12280, 2021b.

Chen, M., Wu, K., Ni, B., Peng, H., Liu, B., Fu, J., Chao, H., and Ling, H. Searching the search space of vision transformer. *Advances in Neural Information Processing Systems*, 34:8714–8726, 2021c.

Choi, J., Wang, Z., Venkataramani, S., Chuang, P. I.-J., Srinivasan, V., and Gopalakrishnan, K. Pact: Parameterized clipping activation for quantized neural networks. *arXiv preprint arXiv:1805.06085*, 2018.

Ding, Y., Qin, H., Yan, Q., Chai, Z., Liu, J., Wei, X., and Liu, X. Towards accurate post-training quantization for vision transformer. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 5380–5388, 2022.

Dong, P., Sun, M., Lu, A., Xie, Y., Liu, K., Kong, Z., Meng, X., Li, Z., Lin, X., Fang, Z., et al. Heatvit: Hardware-efficient adaptive token pruning for vision transformers. In *2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*, pp. 442–455. IEEE, 2023.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Esser, S. K., McKinstry, J. L., Bablani, D., Appuswamy, R., and Modha, D. S. Learned step size quantization. *arXiv preprint arXiv:1902.08153*, 2019.

Frumkin, N., Gope, D., and Marculescu, D. Jumping through local minima: Quantization in the loss landscape of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16978–16988, 2023.

Gong, R., Liu, X., Jiang, S., Li, T., Hu, P., Lin, J., Yu, F., and Yan, J. Differentiable soft quantization: Bridging full-precision and low-bit neural networks. In *International Conference on Computer Vision (ICCV)*, pp. 4852–4861, 2019.

He, Y., Lou, Z., Zhang, L., Liu, J., Wu, W., Zhou, H., and Zhuang, B. Bivit: Extremely compressed binary vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5651–5663, 2023.

Hubara, I., Nahshan, Y., Hanani, Y., Banner, R., and Soudry, D. Accurate post training quantization with small calibration sets. In *International Conference on Machine Learning*, pp. 4466–4475. PMLR, 2021.

Jeon, Y., Lee, C., Cho, E., and Ro, Y. Mr. biq: Post-training non-uniform quantization based on minimizing the reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12329–12338, 2022.

Krishnamoorthi, R. Quantizing deep convolutional networks for efficient inference: A whitepaper. *arXiv preprint arXiv:1806.08342*, 2018.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Le, P.-H. C. and Li, X. Binaryvit: Pushing binary vision transformers towards convolutional models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4664–4673, 2023.

Li, C., Tang, T., Wang, G., Peng, J., Wang, B., Liang, X., and Chang, X. Bossnas: Exploring hybrid cnn-transformers with block-wisely self-supervised neural architecture search. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12281–12291, 2021a.

Li, Y., Gong, R., Tan, X., Yang, Y., Hu, P., Zhang, Q., Yu, F., Wang, W., and Gu, S. {BRECQ}: Pushing the limit of post-training quantization by block reconstruction. In *International Conference on Learning Representations (ICLR)*, 2021b.

Li, Y., Xu, S., Zhang, B., Cao, X., Gao, P., and Guo, G. Q-vit: Accurate and fully quantized low-bit vision transformer. *Advances in Neural Information Processing Systems*, 35:34451–34463, 2022a.

Li, Z. and Gu, Q. I-vit: Integer-only quantization for efficient vision transformer inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17065–17075, 2023.

Li, Z., Ma, L., Chen, M., Xiao, J., and Gu, Q. Patch similarity aware data-free quantization for vision transformers. In *European Conference on Computer Vision*, pp. 154–170. Springer, 2022b.

Li, Z., Xiao, J., Yang, L., and Gu, Q. Repq-vit: Scale reparameterization for post-training quantization of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17227–17236, 2023.

Lin, C., Peng, B., Li, Z., Tan, W., Ren, Y., Xiao, J., and Pu, S. Bit-shrinking: Limiting instantaneous sharpness for improving post-training quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16196–16205, 2023a.

Lin, J., Tang, J., Tang, H., Yang, S., Dang, X., and Han, S. Awq: Activation-aware weight quantization for llm compression and acceleration. *arXiv preprint arXiv:2306.00978*, 2023b.

Lin, Y., Zhang, T., Sun, P., Li, Z., and Zhou, S. Fq-vit: Post-training quantization for fully quantized vision transformer. *arXiv preprint arXiv:2111.13824*, 2021.

Liu, J., Niu, L., Yuan, Z., Yang, D., Wang, X., and Liu, W. Pd-quant: Post-training quantization based on prediction difference metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24427–24437, 2023a.

Liu, Y., Yang, H., Dong, Z., Keutzer, K., Du, L., and Zhang, S. Noisyquant: Noisy bias-enhanced post-training activation quantization for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20321–20330, 2023b.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021a.

Liu, Z., Shen, Z., Li, S., Helwegen, K., Huang, D., and Cheng, K.-T. How do adam and training strategies help bnns optimization. In *International conference on machine learning*, pp. 6936–6946. PMLR, 2021b.

Liu, Z., Wang, Y., Han, K., Zhang, W., Ma, S., and Gao, W. Post-training quantization for vision transformer. *Advances in Neural Information Processing Systems*, 34: 28092–28103, 2021c.

Ma, Y., Li, H., Zheng, X., Xiao, X., Wang, R., Wen, S., Pan, X., Chao, F., and Ji, R. Solving oscillation problem in post-training quantization through a theoretical perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7950–7959, 2023.

Nagel, M., Amjad, R. A., Van Baalen, M., Louizos, C., and Blankevoort, T. Up or down? adaptive rounding for post-training quantization. In *International Conference on Machine Learning*, pp. 7197–7206. PMLR, 2020.

Shao, W., Chen, M., Zhang, Z., Xu, P., Zhao, L., Li, Z., Zhang, K., Gao, P., Qiao, Y., and Luo, P. Omniquant: Omnidirectionally calibrated quantization for large language models. *arXiv preprint arXiv:2308.13137*, 2023.

Sun, M., Ma, H., Kang, G., Jiang, Y., Chen, T., Ma, X., Wang, Z., and Wang, Y. Vaqf: Fully automatic software-hardware co-design framework for low-bit vision transformer. *arXiv preprint arXiv:2201.06618*, 2022.

Tang, Y., Han, K., Wang, Y., Xu, C., Guo, J., Xu, C., and Tao, D. Patch slimming for efficient vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12165–12174, 2022.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

wang, c., Zheng, D., Liu, Y., and Li, L. Leveraging interlayer dependency for post -training quantization. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 6666–6679. Curran Associates, Inc., 2022.

Wang, N., Liu, C.-C. C., Venkataramani, S., Sen, S., Chen, C.-Y., El Maghraoui, K., Srinivasan, V. V., and Chang, L. Deep compression of pre-trained transformer models. *Advances in Neural Information Processing Systems*, 35: 14140–14154, 2022a.

Wang, Z., Wang, C., Xu, X., Zhou, J., and Lu, J. Quantformer: Learning extremely low-precision vision transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022b.

Wei, X., Gong, R., Li, Y., Liu, X., and Yu, F. Qdrop: Randomly dropping quantization for extremely low-bit post-training quantization. *arXiv preprint arXiv:2203.05740*, 2022a.

Wei, X., Zhang, Y., Zhang, X., Gong, R., Zhang, S., Zhang, Q., Yu, F., and Liu, X. Outlier suppression: Pushing the limit of low-bit transformer language models. In *Advances in Neural Information Processing Systems*, 2022b.

Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., and Han, S. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning*, 2023.

Xu, S., Li, Y., Lin, M., Gao, P., Guo, G., Lü, J., and Zhang, B. Q-detr: An efficient low-bit quantized detection transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3842–3851, 2023.

Yang, H., Yin, H., Molchanov, P., Li, H., and Kautz, J. Nvit: Vision transformer compression and parameter redistribution. 2021.

Yang, H., Yin, H., Shen, M., Molchanov, P., Li, H., and Kautz, J. Global vision transformer pruning with hessian-aware saliency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18547–18557, 2023.

Yu, L. and Xiang, W. X-pruner: explainable pruning for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24355–24363, 2023.

Yuan, Z., Xue, C., Chen, Y., Wu, Q., and Sun, G. Ptq4vit: Post-training quantization for vision transformers with twin uniform quantization. In *European Conference on Computer Vision*, pp. 191–207. Springer, 2022.

A. Proof of Theorem 1

Theorem A.1. *Assuming that the two modules satisfy the transformer-like topological homogeneity, for the same input activation, the coarser granularity module corresponds to a smaller quantization error. Formally,*

$$
\begin{aligned}
&\mathcal{L}(X_i, W_{i+1}, \cdots, W_{i+n}) \\
&\leq \mathcal{L}(X_i, W_{i+1}^{'}, \cdots, W_{i+h}^{'}, \cdots, W_{i+n}).
\end{aligned} \tag{6}
$$

Where $\{W_{i+1}^{'}, \cdots, W_{i+h}^{'}\}$ *make* $\mathcal{L}(X_i, W_{i+1}^{'}, \cdots, W_{i+h}^{'})$ *optimal.*

Proof. When the loss function $\mathcal{L}$ is convex, the local optimal solution of $\mathcal{L}$ is part of the global optimal solution. We define the following optimization problem.

$$
\underset{W_{i+1},\cdots,W_{i+n}}{\operatorname{argmin}} \quad \mathcal{L}(X_i, W_{i+1}, \cdots, W_{i+n}). \tag{7}
$$

Assume that the optimal solution to the above optimization problem is $\{W_{i+1}^{*}, \cdots, W_{i+n}^{*}\}$, and if $\mathcal{L}$ is a convex function, then $\{W_{i+1}^{'}, \cdots, W_{i+h}^{'}\} \in \{W_{i+1}^{*}, \cdots, W_{i+n}^{*}\}$. where $\{W_{i+1}^{'}, \cdots, W_{i+h}^{'}\}$ make $\mathcal{L}(X_i, W_{i+1}^{'}, \cdots, W_{i+h}^{'})$ optimal. However, the quantization loss landscape is extremely nonconvex due to the derivatives of the approximate rounding operation (Liu et al., 2021b; Frumkin et al., 2023). Therefore, there exists a parameter set $\{W_{i+1}^{\circ}, \cdots, W_{i+n}^{\circ}\}$ such that,

$$
\begin{aligned}
&\mathcal{L}(X_i, W_{i+1}^{\circ}, \cdots, W_{i+n}^{\circ}) \\
&\leq \mathcal{L}(X_i, W_{i+1}^{'}, \cdots, W_{i+h}^{'}, \cdots, W_{i+n})
\end{aligned} \tag{8}
$$

when $\mathcal{L}(X_i, W_{i+1}^{'}, \cdots, W_{i+h}^{'}) \leq \mathcal{L}(X_i, W_{i+1}^{\circ}, \cdots, W_{i+h}^{\circ})$. We take $\{W_{i+1}, \cdots, W_{i+n}\}$ to be $\{W_{i+1}^{\circ}, \cdots, W_{i+n}^{\circ}\}$ and the theorem is proved.

□

Algorithm 1 Granularity and Optimization

```
Input: ViT blocks $B$, model input $X$, model weight $W$
Output: quantized weight $\hat{W}$
Initialize $DSFlag = False$.
for $i = 1$ to $B$ do
  if $i$ is downsample block then
    $DSFlag = True$
  end if
end for
if $DSFlag == True$ then
  for $i = 1$ to $3B$ do {finer granularity}
    $\underset{\hat{W}}{\operatorname{argmin}}\, \mathcal{L}(X_{i-1}, W_i)$
  end for
else
  for $i = 1$ to $B/n$ do {coarser granularity}
    $\underset{\hat{W}}{\operatorname{argmin}}\, \mathcal{L}(X_{i-1}, W_i, W_{i+1}, \cdots, W_{i+n})$
  end for
end if
```

B. In-stage Search of Swin

Figure 6 illustrates the relationships between final block loss and classifier loss in the Swin model, constrained by transformer-like topological homogeneity. This model's coarser reconstruction granularity, achieved by in-stage sampling restriction, aligns with the quantization performance observed in the ViT and DeiT models, indicating improved quantization outcomes.

C. Gradient Formula Derivation

When the reconstruction module traverses the h-th and $(h+1)$-th stages of Swin, the gradient of the parameter at stage h is defined as follows:

$$
\begin{aligned}
&\frac{\partial \mathcal{L}(X_{i-1}^{h+1}, W_i^{h+1})}{\partial W_j^h} \\
=&\frac{\partial \mathcal{L}(X_{i-1}^{h+1}, W_i^{h+1})}{\partial X_{i-1}^{h+1}} \frac{\partial X_{i-1}^{h+1}}{\partial W_j^h}, \qquad (9) \\
=&\frac{\partial \mathbb{E}\left[\left\| f_i^{h+1}(X_{i-1}^{h+1}, W_i^{h+1}) - f_i^{h+1}(\hat{X}_{i-1}^{h+1}, \hat{W}_i^{h+1})\right\|_F^2\right]}{\partial X_{i-1}^{h+1}} \\
&\times \frac{\partial f_{i-1}^{h+1}(X_{i-2}^{h+1}, W_{i-1}^{h+1})}{\partial X_{i-2}^{h+1}} \frac{\partial X_{i-2}^{h+1}}{\partial W_j^h}, \qquad (10) \\
=&\mathbb{E}\Bigg[2\left\| f_i^{h+1}(X_{i-1}^{h+1}, W_i^{h+1}) - f_i^{h+1}(\hat{X}_{i-1}^{h+1}, \hat{W}_i^{h+1})\right\|_F \\
&\times \left(f_i^{'h+1}(X_{i-1}^{h+1}, W_i^{h+1}) - f_i^{'h+1}(\hat{X}_{i-1}^{h+1}, \hat{W}_i^{h+1})\right) \\
&\times \prod_{a=1}^{i-1} f_a^{'h+1}(X_{a-1}^{h+1}, W_a^{h+1}) \frac{\partial X_0^{h+1}}{\partial W_j^h}\Bigg], \qquad (11) \\
=&\mathbb{E}\Bigg[2\left\| f_i^{h+1}(X_{i-1}^{h+1}, W_i^{h+1}) - f_i^{h+1}(\hat{X}_{i-1}^{h+1}, \hat{W}_i^{h+1})\right\|_F \\
&\times \left(f_i^{'h+1}(X_{i-1}^{h+1}, W_i^{h+1}) - f_i^{'h+1}(\hat{X}_{i-1}^{h+1}, \hat{W}_i^{h+1})\right) \\
&\times \prod_{a=1}^{i-1} f_a^{'h+1}(X_{a-1}^{h+1}, W_a^{h+1}) \frac{\partial f_{DS}(X_l^h, W_{DS})}{\partial X_l^h} \frac{\partial X_l^h}{\partial W_j^h}\Bigg], \\
&\qquad (12) \\
=&\mathbb{E}\Bigg[2\left\| f_i^{h+1}(X_{i-1}^{h+1}, W_i^{h+1}) - f_i^{h+1}(\hat{X}_{i-1}^{h+1}, \hat{W}_i^{h+1})\right\|_F \\
&\times \left(f_i^{'h+1}(X_{i-1}^{h+1}, W_i^{h+1}) - f_i^{'h+1}(\hat{X}_{i-1}^{h+1}, \hat{W}_i^{h+1})\right)
\end{aligned}
$$

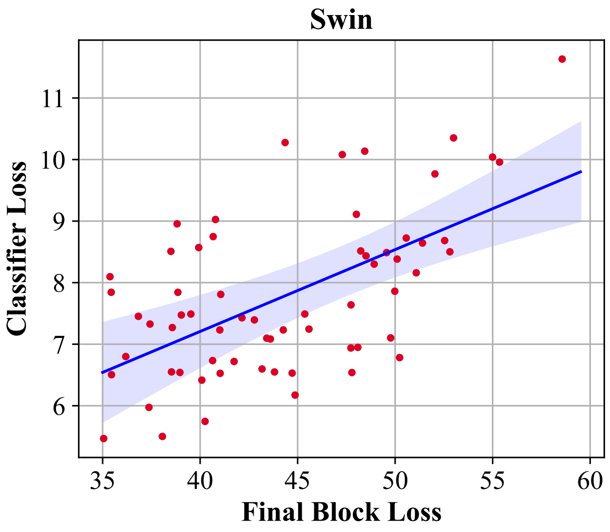


Figure 6. Relationship between final block loss and classifier loss on the Swin model. We maintain an invariant combining scheme for the first, second, and fourth stages, and exclusively sample the transformer block combining scheme in the third stage.

$$\times f'_{DS}(X^h_l, W_{DS}) \times \prod_{a=1}^{i-1} f'^{h+1}_a(X^{h+1}_{a-1}, W^{h+1}_a)$$

$$\times \prod_{b=j}^{l} f'^{h}_b(X^h_{b-1}, W^h_b) \times \frac{\partial f^h_j(X^h_{j-1}, W^h_j)}{\partial W^h_j}\Bigg]. \tag{13}$$

Where subscript represents the transformer block and superscript represents the stage. For example, f'^{h}_b denotes the b-th transformer block of the h-th stage in the Swin, which takes the partial derivative with respect to the activation. x^{h+1}_{i-1} refers to the output activation of the $i-1$)-th transformer block of the $(h+1)$-th stage. The term f_{DS} represents the downsampled block. l denotes the number of transformer blocks in the h-th stage.

Meanwhile, the gradient of the parameters of the module within stage h is,

$$\frac{\partial \mathcal{L}(X^h_{e-1}, W^h_e)}{\partial W^h_j}$$

$$= \frac{\partial \mathcal{L}(X^h_{e-1}, W^h_e)}{\partial X^h_{e-1}} \frac{\partial X^h_{e-1}}{\partial W^h_j}, \tag{14}$$

$$= \frac{\partial \mathbb{E}\left[\left\| f^h_e(X^h_{e-1}, W^h_e) - f^h_e(\hat{X}^h_{e-1}, \hat{W}^h_e)\right\|^2_F\right]}{\partial X^h_{e-1}}$$

$$\times \frac{\partial f^h_{e-1}(X^h_{e-2}, W^h_{e-1})}{\partial X^h_{e-2}} \frac{\partial X^h_{e-2}}{\partial W^h_j}, \tag{15}$$

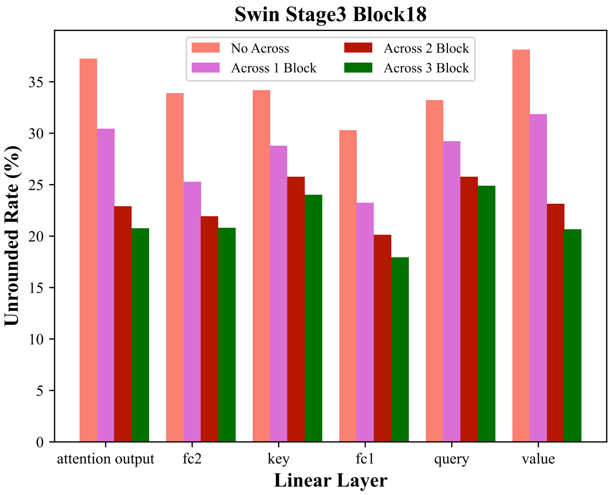


Figure 7. In the Swin model's third stage, we analyze the non-rounding parameter ratios for the final transformer block's six linear layers. "No Across" denotes the absence of cross-stage optimization. "Across 1 Block" refers to joint optimization with one block in the subsequent stage. "Across 2 Blocks" and "Across 3 Blocks" indicate joint optimization with two and three blocks in the next stage, respectively.

$$= \mathbb{E}\Bigg[2\left\| f^h_e(X^h_{e-1}, W^h_e) - f^h_e(\hat{X}^h_{e-1}, \hat{W}^h_e)\right\|_F$$

$$\times \left(f'^{h}_e(X^h_{e-1}, W^h_e) - f'^{h}_e(\hat{X}^h_{e-1}, \hat{W}^h_e)\right)$$

$$\times \prod_{b=j}^{e-1} f'^{h}_b(X^h_{b-1}, W^h_b) \times \frac{\partial f^h_j(X^h_{j-1}, W^h_j)}{\partial W^h_j}\Bigg]. \tag{16}$$

Where e is the e-th transformer block of the h-th stage and $e <= l$.

D. Insufficient Optimization Across Stages

Figure 8 illustrates that the non-stage crossing optimization scheme results in the highest non-rounding parameter ratio across all linear layers. This suggests that parameters in non-crossing stage scenarios are more likely to deviate from initial rounding values, indicating more effective optimization. In contrast, cross-stage optimization scenarios tend to retain initialized rounding values, implying suboptimal optimization. Moreover, as joint optimization block numbers across stages increase, the effectiveness of parameter optimization in earlier stages decreases, evident in the reduced ratio of non-rounded parameters with more combined blocks. Equations (13) and (16) demonstrate how loss gradients at different stages affect parameter optimization. Algorithm 1 outlines the process for determining reconstruction granularity and quantization optimization.

Table 4. Comparison of perplexity effects under the w2a16g128 quantization configuration on Llama1&2.

Method	**LLaMA-7B**		**LLaMA2-7B**	
	WikiText2	**C4**	**WikiText2**	**C4**
FP	5.68	7.08	5.47	6.97
RTN	1.9e3	1.0e3	4.2e3	4.9e3
GPTQ	44.01	27.71	36.77	33.70
AWQ	2.6e5	1.9e5	2.2e5	1.7e5
OmniQuant	9.72	12.97	11.06	15.02
Ours+OmniQuant	8.80	12.03	9.51	13.13

E. LLM Results

We integrate our method with OmniQuant (Shao et al., 2023) on LLaMA1&2 and pursue coarse granularity quantization as per Rule 1, choosing a 4-block granularity to evenly split the 32 hidden blocks. Without loss of generality, we evaluate the w2a16g128 quantization scheme, with detailed results presented below. Our approach uniquely explores granularity in PTQ optimization for language models, setting it apart from existing PTQ methods like SmoothQuant (Xiao et al., 2023), AWQ (Lin et al., 2023b), Outlier Suppression (Wei et al., 2022b), and OmniQuant (Xiao et al., 2023), which mainly focus on equivalent transformations to minimize quantization errors.

F. Experimental Settings

Our approach aligns with the hyperparameter settings used in Adaround (Nagel et al., 2020), BRECQ (Li et al., 2021b), and QDrop (Wei et al., 2022a). We use 16 batches of 64 samples each from the training set for calibration. The learning rates are set at 1e-3 for the rounding parameter and 4e-5 for the quantization scale of the activation layer. The rounding loss rate is set at 0.1, with 20,000 iterations per optimization block. The activation value drop probability is 50%. We gradually reduce the power β of the progressive soft function from 20 to 2. For activation calibration, we use the EMA-mean-square-error method, and for weights, we employ min-max calibration.

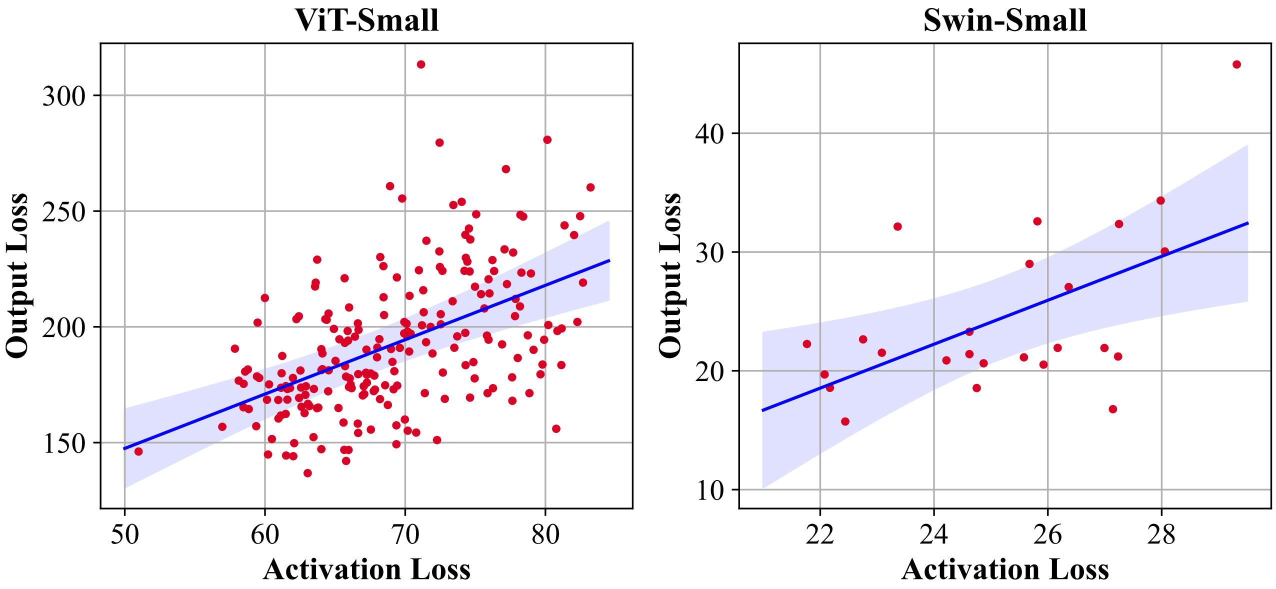


Figure 8. The relationship between initial input activation loss and post-optimization output loss. By randomly sampling initial activations for ViT-Small and Swin-Small's last block and optimizing with a 4/4 bit quantization setup, we observe a positive correlation between the magnitude of initial loss and final optimization loss.