Quantformer: Learning Extremely Low-precision Vision Transformers (Supplementary Material)

Ziwei Wang, *Student Member, IEEE,* Changyuan Wang, Xiuwei Xu, *Student Member, IEEE,* Jie Zhou, *Senior Member, IEEE,* and Jiwen Lu, *Senior Member, IEEE*

APPENDIX A: IMPLEMENTATION DETAILS FOR MIXED-PRECISION QUANTIZATION IN SECTION 4.3.1

In order to further enhance the performance of Quantformer, we integrated the presented techniques with mixed-precision quantization methods including HAQ [6] and EdMIPS [1]. Mixed-precision quantization selects the bitwidth for each layer according to the informativeness to achieve more optimal accuracy-complexity trade-offs.

For the implementation of HAQ on vision transformers, we applied deep deterministic policy gradient (DDPG) [3] to predict the optimal bitwidth of each fully-connected layer in vision transformers, where the input was set as the state and the action of last step. The critic networks consisted of two fully-connected layers with the hidden size as 400, and the actor networks included two fully-connected layers with hidden size as 300 and extra two hidden vectors. The output of actor networks was fed forward into sigmoid function to be ranged into $[0, 1]$. The optimization of the DDPG used the AdamW optimizer [2] with the learning rate $1e$-4 for the actor networks and $1e$-3 for the critic networks respectively. The weight noise during the exploration process was generated from truncated normal distribution ranged in $[0, 1]$, where the standard error was initialized as 0.5 and decayed by 0.99 after each epoch. The data for exploration was randomly selected from 100 classes of ImageNet, and the finetuning epoch number of quantized models for reward acquisition during the search stage was set as one.

The learning rate of AdamW optimizer applied in EdMIPS was initialized as $1e$-4 and $1e$-5 for the network parameters and the architecture parameters respectively, where all architecture parameters were equal for the initialization. The supernet was trained by 50 epochs with the cosine annealing strategy for learning rate decay, where the ending learning rates were $100\times$ less than the initial ones. After the search process, the final model was obtained by discretizing the supernet with the branch in largest importance weights.

The mixed-precision vision transformers acquired by HAQ and EdMIPS were finetuned by 120 epochs with the AdamW optimizer [2]. The learning rate was flexibly initialized according to the model complexity, where more lightweight transformers leveraged smaller initialized learning rate and vice versa. The cosine annealing strategy for learning rate decay was also applied with the ending learning rate $1e$-6. For the combination of mixed-precision quantization and our Quantformer, we also followed the details described in Section 4.1 of the manuscript to implement self-attention rank preservation and group-wise quantization. The batchsize was set to 512 for all experiments in the integration of mixed-precision quantization methods and Quantformer.

APPENDIX B: COMPARISON WITH RANKING-AWARE QUANTIZATION IN [4]

Ranking-aware quantization for self-attention [4] considers the self-attention rank consistency between the quantized and full-precision vision transformers, and they apply hinge loss as the optimization objective:

$$L = \sum_{k=1}^{h} \sum_{i=1}^{w-1} \sum_{j=i+1}^{w} \Phi((\hat{\boldsymbol{A}_{ki}} - \hat{\boldsymbol{A}_{kj}}) \cdot sign(\boldsymbol{A}_{ki} - \boldsymbol{A}_{kj})) \quad (1)$$

where $\Phi(p) = (\theta - p)_+$ is the hinge function with the hyperparameter θ, and (h, w) is the size of the self-attention matrix. $\hat{\boldsymbol{A}_{ki}}$ and $\boldsymbol{A}_{ki}$ are the elements in the k_{th} row and i_{th} column of the quantized and full-precision self-attention respectively. The loss achieves zero only when the self-attention element pairs are in correct order and differed by a margin. Although the hinge loss can benefit self-attention rank consistency preservation, it ignores the entropy variance in quantized self-attention caused by capacity difference. In Figure 1, we visualize the self-attention element distribution of quantized vision transformers optimized by

- *Ziwei Wang, Changyuan Wang, Xiuwei Xu, Jie Zhou, and Jiwen Lu are with the Beijing National Research Center for Information Science and Technology (BNRist), Department of Automation, Tsinghua University, Beijing 100084, China. E-mail: wang-zw18@mails.tsinghua.edu.cn, 201811210202@mail.bnu.edu.cn, xxw21@mails.tsinghua.edu.cn, jzhou@tsinghua.edu.cn, lujiwen@tsinghua.edu.cn.*
- *Corresponding author: Jiwen Lu*

- *Code: https://github.com/ZiweiWangTHU/Quantformer.git.*

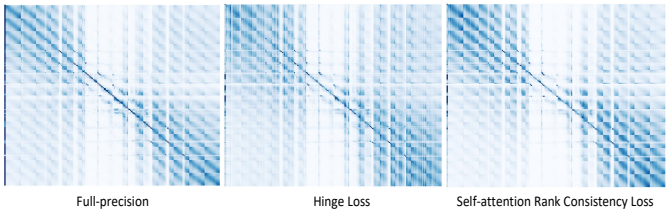

Figure 1. The visualization for self-attention for the 2_{nd} layer in DeiT-T of (a) full-precision models, (b) 4-bit models optimized by [4], (c) 4-bit models optimized by our method.

Table 1
The accuracy and the training cost of quantized DeiT-T models optimized by hinge loss and (7) in the manuscript.

Precision	Methods	Top-1	Top5	Training Cost
4-bit	Hinge loss	68.1	88.7	120 GPU hours
	Quantformer	69.9	89.7	28 GPU hours
3-bit	Hinge loss	61.5	84.6	120 GPU hours
	Quantformer	65.2	87.0	28 GPU hours
2-bit	Hinge loss	57.9	82.0	120 GPU hours
	Quantformer	60.7	84.0	28 GPU hours

the hinge loss and (7) in the manuscript, where the self-attention in the 2_{nd} layer of 4-bit DeiT-T is leveraged for demonstration. Each element pair with the correct relative order is equally encouraged in [4] to achieve the difference over the margin, which results in excessively high entropy for quantized self-attention with capacity insufficiency. Since the computational complexity of the hinge loss and our capacity-aware self-attention rank consistency loss is respectively $O(hw^2)$ and $O(hw)$, calculating the exact value of the hinge loss is much more complicated than (7) in the manuscript, which causes higher training cost in quantization-aware training. Table 1 illustrates the accuracy and the training cost of the quantized DeiT-T optimized by the hinge loss in [4] and (7) in the manuscript, where our loss achieves higher accuracy due to the capacity-aware self-attention distribution and requires lower training cost.

C. Technical Soundness of the Self-attention Rank Consistency Loss in (7)

The goal of (6) in the manuscript contains (a) keeping the self-attention rank consistency between the quantized and full-precision transformers, and (b) enabling the self-attention distribution to be adjusted according to the network capacity. As (6) in the manuscript cannot be directly optimized via back-propagation, we present the surrogate loss shown by (7) in the manuscript to achieve the self-attention rank consistency preservation with capacity-aware distribution. The intuition is that by changing the power of full-precision self-attention, the distribution concentration can be adjusted without modifying the rank, so that minimizing (7) in the manuscript can simultaneously realize the above two goals. In order to verify the technical soundness of (7) in the manuscript, we have provided the theoretical proof for the strong correlation between the (6) and (7) in the manuscript, shown the model statistics of the self-attention rank in the quantized and full-precision vision transformers and conducted ablation studies that leverage different alternatives to optimize the original loss.

Theoretical proof: Let us assume the query $\boldsymbol{Q}$ and key $\boldsymbol{K}$ satisfying the Gaussian distribution, and the variable $\frac{\boldsymbol{QK}}{\sqrt{d}}$ also meets the Gaussian distribution. Denoting the elements in the matrix $\frac{\boldsymbol{QK}}{\sqrt{d}}$ as x, the probability distribution function (PDF) of the variable x can be written as follows with the mean μ and the standard deviation σ:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2}$$

The subsequent softmax function for self-attention calculation with row-wise normalization can be represented in the following:

$$a_i = \frac{e^{x_i}}{\sum_{k=1}^{n} e^{x_k}} \tag{3}$$

where a_i and x_i represent the i_{th} element in the given row of self-attention and $\frac{\boldsymbol{QK}}{\sqrt{d}}$ respectively. As n is usually large (192 for DeiT-T), we can regard that the dominator in (3) is a constant S and independent of x_i because e^{x_i} only makes little contribution to S. For the softmax function, shifting all x variables by a constant distance does not affect the value of a, so that we shift all $\frac{\boldsymbol{QK}}{\sqrt{d}}$ by μ to simplify the distribution. The distribution of the elements in self-attention is written as follows, where we omit the subscript i for simplicity:

$$p(a) = \frac{1}{\sqrt{2\pi}\sigma a} e^{-\frac{(log(Sa))^2}{2\sigma^2}} \tag{4}$$

where $log(x)$ means the natural logarithm. Since the cumulative probability function (CDF) for a can approximately demonstrate the ranking of the self-attention element in the same row, the self-attention rank preservation objective shown by (6) in the manuscript can be represented as follows:

$$\begin{aligned} J_{src} &= || \int_{-\infty}^{a_q} p(a_q) da_q - \int_{-\infty}^{a_r} p(a_r) da_r || \\ &= ||\Phi(\frac{\log(a_q) - \log(\frac{1}{S_q})}{\sigma_q}) - \Phi(\frac{\log(a_r) - \log(\frac{1}{S_r})}{\sigma_r})|| \end{aligned} \tag{5}$$

where $\Phi(x)$ is the CDF of standard normal distribution for variable x, and the subscript q and r mean corresponding variables in quantized and full-precision vision transformers respectively. The optimization objective can be equivalently formulated as follows:

$$\begin{aligned} \min J_{src} &\iff \min || \frac{\log(a_q) - \log(\frac{1}{S_q})}{\sigma_q} - \frac{\log(a_r) - \log(\frac{1}{S_r})}{\sigma_r} || \\ &\iff \min || a_q - a_r^{\sigma_q/\sigma_r} \cdot \frac{S_r^{\sigma_q/\sigma_r}}{S_q} || \end{aligned} \tag{6}$$

According to the constraint in (6) of the manuscript and the definition of self-attention a in (3), we can obtain the following equation:

$$\log S_q - \frac{\sum_{i=1}^{n} x_{i,q} e^{x_{i,q}}}{\sum_{i=1}^{n} e^{x_{i,q}}} = C_l (\log S_r - \frac{\sum_{i=1}^{n} x_{i,r} e^{x_{i,r}}}{\sum_{i=1}^{n} e^{x_{i,r}}}) \tag{7}$$

Denoting the minimum and maximum value of the sequence $\{x_i\}$ as x_{min} and x_{max} respectively, $\frac{\sum_{i=1}^{n} x_i e^{x_i}}{\sum_{i=1}^{n} e^{x_i}} \in [x_{min}, x_{max}]$ is much smaller than S because the number of self-attention elements in each row denoted as n is usually

large. The following equation can be accurate approximation for (7):

$$\log S_q = C_l \log S_r \tag{8}$$

We can also acquire the continuous form of the constraint in (6) of the manuscript as follows:

$$\int_0^1 a_q \log a_q p(a_q) da_q = C_l \int_0^1 a_r \log a_r p(a_r) da_r \tag{9}$$

To calculate the integral in (9), the PDF of a can be approximated by the power function in the interval $a \in [\frac{1}{S}, b]$ with the constant b:

$$p^*(a) = \mathbb{I}(a \in [\frac{1}{S}, b]) \cdot \frac{(k-1)}{S^{k-1} - \frac{1}{b^{k-1}}} \cdot \frac{1}{a^k} \tag{10}$$

For the value of $a \log a$, it approaches zero in the interval $a \in (0, \frac{1}{S})$ since S is very large. Since $p(a)$ decays very quickly and b does not approach zero, the PDF of a also approaches zero for $a \in (b, 1)$. Therefore, the expectation of $a \log a$ for the continuous self-attention distribution can be rewritten as follows:

$$\begin{aligned}
&\int_0^1 a \log a p(a) da \approx \int_{\frac{1}{S}}^b a \log a p(a) da \approx \int_{\frac{1}{S}}^b a \log a p^*(a) da \\
&= -\frac{(k-1)(k-2)}{b^{k-2} S^{k-1} - \frac{1}{b}} \{\log b + (Sb)^{k-2} \log S + \frac{1}{k-2} \cdot [1 - \\
&\quad (Sb)^{k-2}]\}
\end{aligned} \tag{11}$$

Since S is much larger than b (because the number of self-attention elements in each row denoted as n is large) and k is larger than 2 in the PDF approximation, the following approximation holds:

$$\int_0^1 a \log a p(a) da \approx -\frac{(k-1)(k-2)}{S^k - \frac{S}{b^{k-1}}} S^{k-1} \log S \tag{12}$$

We assign the value of b with $b = (\frac{S}{S^k - d_0})$ where d_0 is a large constant in the same order with S^k. The relationship between quantized and full-precision self-attention in the continuous form can be written as:

$$\begin{aligned}
&\frac{k_q^2 - 3k_q + 2}{d_0} S_q^{k_q - 1} \log S_q = C_l \frac{k_q^2 - 3k_q + 2}{d_0} S_r^{k_r - 1} \log S_r \\
&\iff \frac{k_q^2 - 3k_q + 2}{d_0} S_q^{k_q - 1} = \frac{k_q^2 - 3k_q + 2}{d_0} S_r^{k_r - 1} \\
&\iff \frac{k_q^2 - 3k_q + 2}{d_0} S_q^{k_q - 1} = \frac{k_q^2 - 3k_q + 2}{d_0} S_q^{C_l(k_r - 1)}
\end{aligned} \tag{13}$$

Since (13) holds for any self-attention, so the following equation always holds:

$$k_q - 1 = C_l(k_r - 1) \tag{14}$$

The approximated distribution of the self-attention should be equal with the true distribution at the point $a = \frac{1}{S}$ for accurate approximation, we have the following equation:

$$\begin{aligned}
&p(\frac{1}{S}) = p^*(\frac{1}{S}) \\
&\iff \frac{S}{\sqrt{2\pi\sigma}} = \frac{(k-1)}{S^{k-1} - \frac{1}{b^{k-1}}} * S^k \\
&\underset{S >> \frac{1}{b}}{\iff} \frac{1}{\sqrt{2\pi\sigma}} = k - 1
\end{aligned} \tag{15}$$

Table 2
The accuracy and the self-attention rank difference (SARD) for top-100 pixels in the full-precision self-attention for different bitwidths and optimization methods in DeiT-T.

Precision	Methods	Top-1	Top5	SARD
32-bit	-	72.2	91.1	-
4-bit	PACT	65.6	87.2	27.81
	Hinge loss	68.1	88.7	10.41
	Quantformer	69.9	89.7	9.70
3-bit	PACT	60.6	83.9	45.60
	Hinge loss	61.5	84.6	15.61
	Quantformer	65.2	87.0	14.10

Combining (14) and (15), we acquire the relationship between σ_q and σ_r in (6):

$$\sigma_q = \frac{\sigma_r}{C_l} \tag{16}$$

For x satisfying the distribution shown in (2), the new variable $y = \frac{x}{C_l}$ also meets the Gaussian distribution with the mean $\frac{\mu}{C_l}$ and standard deviation $\frac{\sigma}{C_l}$. For elements in $\frac{\boldsymbol{QK}}{\sqrt{d}}$ in the full-precision vision transformers denoted as $x_{i,r}$, the standard deviation of $y_{i,r} = \frac{x_{i,r}}{C_l}$ equals to $\frac{\sigma_r}{C_l} = \sigma_q$. Since the mean of $x_{i,r}$ and $x_{i,q}$ both approaches zero in practical distribution, the distribution of $y_{i,r}$ is the same as $x_{i,q}$. Therefore, we have the following equation for large n:

$$\sum_{i=1}^n e^{x_{i,q}} = \sum_{i=1}^n e^{\frac{x_{i,r}}{C_l}} \tag{17}$$

The coefficient in the objective is written as follows:

$$\frac{S_r^{\sigma_q/\sigma_r}}{S_q} = \frac{(\sum_{i=1}^n e^{x_{i,r}})^{\frac{1}{C_l}}}{\sum_{i=1}^n e^{x_{i,q}}} = \frac{(\sum_{i=1}^n e^{x_{i,r}})^{\frac{1}{C_l}}}{\sum_{i=1}^n e^{\frac{x_{i,r}}{C_l}}} = (\sum_{i=1}^n a_{i,r}^{\frac{1}{C_l}})^{-1} \tag{18}$$

The layer-wise parameter C_l can be evaluated by $C_l = \frac{\sum_{mn} A_{q,mn}^l \log A_{q,mn}^l}{\sum_{mn} A_{r,mn}^l \log A_{r,mn}^l}$ for each layer to avoid large fluctuation, and the objective for self-attention consistency preservation shown by (6) in the manuscript can be rewritten as follows:

$$\min \sum_{i=1}^n ||a_{i,q} - (\sum_{i=1}^n a_{i,r}^{\frac{1}{C_l}})^{-1} \cdot a_{i,r}^{\frac{1}{C_l}}|| \tag{19}$$

which is in the same form as (7) and (8) in the manuscript.

Model statistics: In order to demonstrate that optimizing (7) in the manuscript can preserve the self-attention rank consistency as (6) in the manuscript shows, Table 2 depicts the self-attention rank difference (SARD) for top-100 pixels in the full-precision self-attention for various bitwidth settings and optimization methods, where the top-1 accuracy on ImageNet is also provided for reference. The results indicate our self-attention rank consistency loss can significantly reduce the SARD for vision transformers, and the performance on SARD is comparable with the ranking-aware quantization presented in [4]. Since [4] fails to consider the capacity variance in vision transformers with different bitwidths, the top-1 accuracy drops significantly compared with their full-precision counterparts due to the capacity insufficiency.

We also provide variable statistics to verify the assumption rationality in the theoretical proof. Figure 2(a) shows an

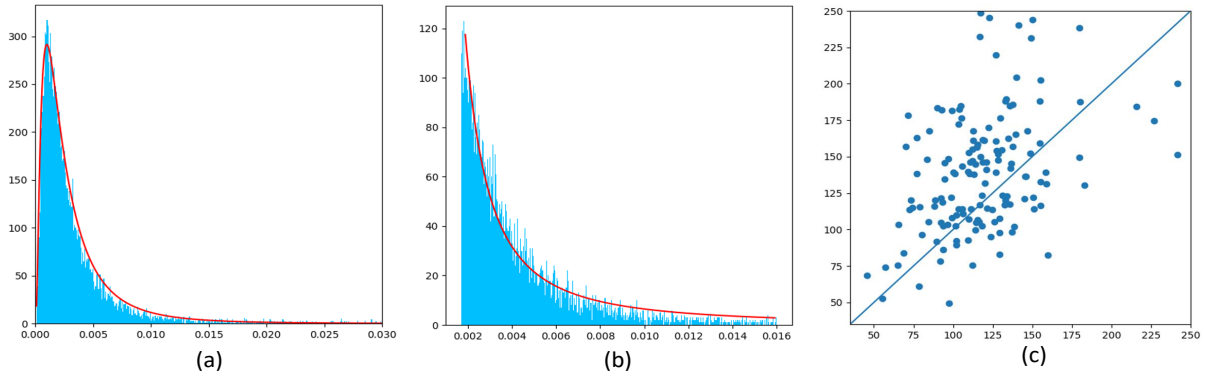

Figure 2. (a) An example of the row-wise self-attention distribution and the distribution shown in (4). We ensemble rows with similar S in the regression to avoid variance caused by sparse sampling. (b) An example of the self-attention and the approximated distribution shown in (10). (c) The value of $\sum_{i=1}^{n} e^{x_{i,q}}$ and corresponding $\sum_{i=1}^{n} e^{\frac{x_{i,r}}{C_l}}$ in (17) for different rows of self-attention.

Table 3
The accuracy of quantized DeiT-T with different bitwidth settings and self-attention rank preservation objectives.

Objective	Form	Bitwidth	Top-1	Top-5
Hinge Loss	-	2	57.9	82.0
		3	61.5	84.6
		4	68.1	88.7
Capacity-aware p_l	$p_l = \sqrt{\frac{E_r}{E_q}}$	2	59.5	83.1
		3	64.0	86.6
		4	69.1	89.5
	$p_l = \exp(\frac{E_r}{E_q})$	2	58.7	82.8
		3	64.3	86.6
		4	69.0	89.2
	$p_l = \frac{E_r}{E_q}$	2	60.7	84.0
		3	65.2	87.0
		4	69.9	89.7
	Random p_l	2	55.3	80.7
		3	59.7	83.1
		4	66.8	88.2

Table 4
The accuracy of quantized DeiT-T with different bitwidth settings and fixed p_l in self-attention rank preservation objectives.

p_l	2-bit		3-bit		4-bit	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
1	57.9	82.3	63.0	85.6	68.5	89.0
2	58.2	82.4	63.1	85.5	68.9	89.2
3	58.3	82.3	63.5	85.9	68.4	89.1
4	58.5	82.5	62.5	85.2	67.8	88.5
5	57.1	82.0	62.4	85.0	67.0	88.2

example of the row-wise self-attention distribution, and the average correlation coefficients for fitting the distribution shown in (4) across all self-attention layers and test images in ImageNet are 0.983 and 0.991 for quantized and full-precision vision transformers, which are relatively high indicating good fitness. Therefore, we conclude that the probability density function (4) can accurately represent the actual distribution of self-attention.

We also fit the self-attention distribution for $a \in [\frac{1}{S}, b]$ with $p^*(a)$ shown in (10), and Figure 2(b) shows several examples with the mean correlation coeficient 0.979. We set S as the value acquired by fitting the true distribution in (4), and regress the value of k by assigning the large constant d_0 with 600 for both quantized and full-precision vision transformers. The average value of b for quantized and full-precision self-attention are 47.9 and 51.2 respectively, so that the assumption that S is much larger than b holds. The average value of $p(a)$ at the point $a = b$ is 0.015 approaching zero, which proves that the integral approximation in (11) holds.

Figure 2(c) shows the value of $\sum_{i=1}^{n} e^{x_{i,q}}$ and $\sum_{i=1}^{n} e^{\frac{x_{i,r}}{C_l}}$ in (17) for different rows of self-attention. We observe that the two terms are strongly correlated for large n.

Comparison with other ranking preservation techniques: To illustrate the effectiveness of capacity-aware self-attention rank preservation loss, we conducted ablation studies by utilizing other rank preservation loss and assigning different loss forms for capacity-aware self-attention rank preservation. For other rank preservation loss, we leverage the widely adopted hinge loss [4], [5], [7] shown in (1). For other loss forms of capacity-aware self-attention rank preservation, we assign p_l with random numbers, various constants and different combination of self-attention entropy. The ablation studies were conducted on ImageNet with the DeiT-T architecture in 2, 3, 4 bits, and the results are reported in Table 3-4.

Although the hinge loss can preserve the rank consistency between the quantized and full-precision self-attention, it ignores the model capacity variance among different bitwidths. As a result, the capacity insufficiency in low-precision networks degrades the performance significantly. Our capacity-aware self-attention rank preservation loss with constant p_l acquires worse performance compared with Quantformer, because the inaccurate capacity estimation leads to capacity insufficiency for networks in low capacity and fails to fully utilize the capacity for networks in high capacity. Meanwhile, randomly p_l underperforms the vanilla quantized vision transformers because the uninformative change of p_l causes convergence problems. For assigning p_l with different combinations of self-attention entropy, the division between the entropy of the full-precision and quantized self-attention achieves the highest accuracy as it is consistent with the objective in (19).

In conclusion, leveraging (7) and (8) in the manuscript as alternatives of (6) in the manuscript for capacity-aware self-attention rank preservation is technically sounded, and

Table 5
The storage cost, computational complexity and the accuracy on ImageNet of different network quantization methods across various vision transformer architectures and bitwidth settings. Param. depicts the number of network parameters and BOPs is the bit-operations, which evaluate the storage and computational cost respectively. W/o SRC and w/o GQ respectively demonstrate Quantformer without self-attention rank consistency loss and without group-wise quantization.

Model	Method	2-bit				3-bit				4-bit			
		Param.	BOPs	Top-1	Top-5	Param.	BOPs	Top-1	Top-5	Param.	BOPs	Top-1	Top-5
DeiT-T	Full-precision	5.11M	1.3T	72.2	91.1	5.11M	1.3T	72.2	91.1	5.11M	1.3T	72.2	91.1
	OCS	0.40M	22.5G	57.2	81.2	0.54M	27.9G	60.4	83.5	0.69M	35.0G	65.7	86.9
	W/o SRC	0.39M	23.2G	59.8	83.7	0.54M	28.5G	63.9	86.2	0.69M	35.8G	68.4	88.6
	W/o GQ	0.39M	22.3G	57.9	82.1	0.53M	27.6G	62.0	84.8	0.69M	34.9G	68.0	88.5
	Quantformer	0.39M	23.2G	60.7	84.0	0.54M	28.5G	65.2	87.0	0.69M	35.8G	69.9	89.7
DeiT-S	Full-precision	22.10M	4.7T	79.9	95.0	22.10M	4.7T	79.9	95.0	22.10M	4.7T	79.9	95.0
	OCS	1.58M	54.2G	60.7	84.0	2.23M	74.9G	72.8	91.0	2.90M	104.5G	75.2	92.7
	W/o SRC	1.55M	54.8G	63.7	86.4	2.22M	75.7G	74.1	92.3	2.90M	105.0G	77.2	93.7
	W/o GQ	1.55M	53.0G	61.9	84.7	2.22M	73.9G	73.0	91.0	2.89M	103.2G	76.9	93.2
	Quantformer	1.55M	54.8G	65.2	87.1	2.22M	75.7G	75.4	92.8	2.90M	105.0G	78.2	94.2
DeiT-B	Full-precision	86.60M	18.0T	81.8	95.6	86.60M	18.0T	81.8	95.6	86.60M	18.0T	81.8	95.6
	OCS	5.79M	140.7G	69.4	89.1	8.48M	226.8G	77.2	93.5	11.23M	345.2G	77.7	93.4
	W/o SRC	5.71M	143.2G	71.5	90.7	8.38M	226.9G	77.9	93.6	11.04M	344.0G	79.0	93.8
	W/o GQ	5.71M	139.5G	71.2	90.2	8.38M	223.2G	77.1	93.7	11.04M	340.3G	78.8	94.0
	Quantformer	5.71M	143.2G	73.8	92.0	8.38M	226.9G	78.3	93.9	11.04M	344.0G	79.7	94.3

performs better than other rank preservation loss and objective forms.

Appendix D. Results Comparing with More Baseline Methods

We conducted experiments to compare our Quantformer with more baseline methods on ImageNet datasets across different vision transformer architectures. The baseline methods include OCS [8], Quantformer without SRC loss, Quantformer without group-wise quantization. The goal of the OCS method [8] is dealing with the outliers that cause large quantization errors. OCS duplicates the channels containing outliers and halves the channel values including weights and activations. The network remains functionally identical while the outliers are moved towards the distribution center. In order to verify the effectiveness of each presented technique, we also leverage Quantformer without SRC loss and Quantformer without group-wise quantization as baselines. Table 5 demonstrates the results, where our Quantformer outperforms the listed baseline methods by a sizable margin. Although OCS [8] and Quantformer shares the same goal for quantization error minimization with slight computation and storage overhead, the quantization policy with shared thresholds and discretization points for patch features distributing diversely across channels still causes sizable clipping and rounding errors. Meanwhile, the capacity-aware self-attention rank preservation and group-wise quantization both make contributions to the top-1 accuracy.

Appendix E. Comparison among Layer-, Channel-, Group-wise Quantization

Layer-wise quantization for features in vision transformers significantly degrades the performance due to the large rounding and clipping errors, while channel-wise quantization strategies add extremely high computation overhead because of the rescaling before accumulation in integer arithmetic. Therefore, we present group-wise quantization to achieve more optimal accuracy-complexity trade-offs via adopting same quantization strategies for feature elements distributed similarly. Table 6 shows the top-1 accuracy on ImageNet, the number of parameters and BOPs of quantized Deit-T, where our Quantformer is more practical in realistic applications.

Table 6
The accuracy of 4-bit Quantformer, where different quantization strategies including layer-, channel-, group-wise quantization are adopted. The architecture is Deit-T. Param. depicts the number of network parameters and BOPs is the bit-operations.

#Groups.	Params.	BOPs	Top-1
Layer-wise	0.69M	34.9G	68.0
Channel-wise	0.84M	109.8G	70.5
Group-wise	0.69M	35.8G	69.9

References

[1] Zhaowei Cai and Nuno Vasconcelos. Rethinking differentiable search for mixed-precision neural networks. In *CVPR*, pages 2349–2358, 2020.

[2] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[3] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[4] Zhenhua Liu, Yunhe Wang, Kai Han, Siwei Ma, and Wen Gao. Post-training quantization for vision transformer. *arXiv preprint arXiv:2106.14156*, 2021.

[5] Mohammad Norouzi, David J Fleet, and Russ R Salakhutdinov. Hamming distance metric learning. *NIPS*, 25, 2012.

[6] Kuan Wang, Zhijian Liu, Yujun Lin, Ji Lin, and Song Han. Haq: Hardware-aware automated quantization with mixed precision. In *CVPR*, pages 8612–8620, 2019.

[7] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 10(2), 2009.

[8] Ritchie Zhao, Yuwei Hu, Jordan Dotzel, Chris De Sa, and Zhiru Zhang. Improving neural network quantization without retraining using outlier channel splitting. In *ICML*, pages 7543–7552, 2019.