Q-ViT: Accurate and Fully Quantized Low-bit Vision Transformer

Anonymous Author(s)
Affiliation
Address
email

Abstract

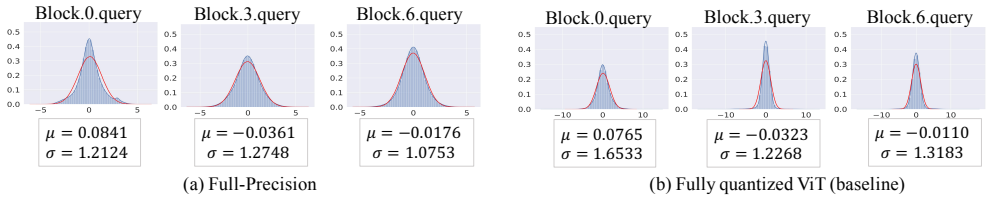
The large pre-trained vision transformers (ViTs) have demonstrated remarkable performance on various visual tasks, but suffer from expensive computational and memory cost problems when deployed on resource-constrained devices. Among the powerful compression approaches, quantization extremely reduces the computation and memory consumption by low-bit parameters and bit-wise operations. However, low-bit ViTs remain largely unexplored and usually suffer from a significant performance drop compared with the real-valued counterparts. In this work, through extensive empirical analysis, we first identify the bottleneck for severe performance drop comes from the information distortion of the low-bit quantized self-attention map. We then develop an information rectification module (IRM) and a distribution guided distillation (DGD) scheme for fully quantized vision transformers (Q-ViT) to effectively eliminate such distortion, leading to a fully quantized ViTs. We evaluate our methods on popular DeiT and Swin backbones. Extensive experimental results show that our method achieves a much better performance than the prior arts. For example, our Q-ViT can theoretically accelerates the ViT-S by 6.14× and achieves about 80.9% Top-1 accuracy, even surpassing the full-precision counterpart by 1.0% on ImageNet dataset. Test code and models are attached on `https://github.com/AnonymousResultsNips/NeurIPs659`

1 Introduction

Inspired by the success in natural language processing (NLP), transformer-based models have shown great power in various computer vision (CV) tasks, such as image classification [4] and object detection [2]. Pre-trained with large-scale data, these models usually have a tremendous number of parameters. For example, there are 632M parameters taking up 2528MB memory usage and 162G FLOPs in the ViT-H model, which is both memory and computation expensive during inference. This limits these models for the deployment on resource-limited platforms. Therefore, compressed transformers are urgently needed for real applications.

Substantial efforts have been made to compress and accelerate neural networks for efficient online inference. Methods include compact network design [9], network pruning [8], low-rank decomposition [3], quantization [20], and knowledge distillation [23]. Quantization is particularly suitable for deployment on AI chips because it reduces the bit-width of network parameters and activations for efficient inference. Prior post-training quantization (PTQ) methods [17, 13] on ViTs directly compute quantized parameters based on pre-trained full-precision models, which constrains the model performance to a sub-optimized level without fine-tuning. Furthermore, quantizing these

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.



(a) Full-Precision (b) Fully quantized ViT (baseline)

Figure 1: The histogram of query values $\mathbf{q}$ (blue shadow) along with the PDF curve (red line) of Gaussian distribution $N(\mu, \sigma^2)$ [19], for 3 selected layers in DeiT-T and 4-bit fully quantized DeiT-T (baseline). μ and σ^2 are the statistical mean and variance of the values.

models based on PTQ methods to ultra-low bits (*e.g.*, 4 bits or lower) is ineffective and suffers from a significant performance reduction.

Differently, quantization-aware training (QAT) [15] methods perform quantization during back propagation and achieve much less performance drop with a higher compression rate generally. QAT is shown to be effective for CNN models [16] for CV tasks. However, QAT methods remain largely unexplored for low-bit quantization of vision transformers. Therefore, we first build a fully quantized ViT baseline, a straightforward yet effective solution based on common techniques. Our study discovers that the performance drop of fully quantized ViT lies in the information distortion among the attention mechanism in the forward process, and the ineffective optimization for eliminating the distribution difference through distillation in the backward propagation. First, the attention mechanism of ViT aims at modeling long-distance dependencies [26, 4]. However, our analysis shows that a direct quantization method leads to the information distortion, *i.e.*, significant distribution variation for the query module between quantized ViT and full-precision counterpart. For example, as shown in Fig. 1, the variance difference is 0.4409 (1.2124 *v.s.* 1.6533) for the first block [1]. This inevitably deteriorates the representation capability of the attention module on capturing the global dependency for the input. Second, the distillation for the fully quantized ViT baseline utilizes distillation token (following [24]) to directly supervise the classification output of the quantized ViT. However, we found that such a simple supervision is ineffective, which is coarse-grained for the large gap between the quantized attention scores and their full-precision counterparts.

To address the aforementioned issues, a fully quantized ViT (Q-ViT) is developed by retaining the distribution of quantized attention modules as that of full-precision counterparts (see the overview in Fig. 2). Accordingly, we propose to modify the distorted distribution over quantized attention modules through an Information Rectification Module (IRM) based on information entropy maximization, in the forward process. While in the backward process, we present a Distribution Guided Distillation (DGD) scheme to eliminate the distribution variation through attention similarity loss between quantized ViT and full-precision counterpart. The contributions of our work include:

- We propose an Information Rectification Module (IRM) based on the information theory to address the information distortion problem. IRM applies quantized representations in the attention module with a maximized information entropy, allowing the quantized model to restore the representation of input images.
- We develope a Distribution Guided Distillation (DGD) scheme to eliminate the distribution mismatch in distillation. DGD takes appropriate activations and utilizes knowledge from the similarity matrices in distillation to perform optimization accurately.
- Our Q-ViT, for the first time, explores a promising way towards accurate and low-bit ViT. Extensive experiments on the ImageNet benchmark show that our Q-ViT outperforms the baseline by a large margin, and achieves comparable performances with the full-precision counterparts.

[1] The Gaussian distribution hypothesis is supported by [19]

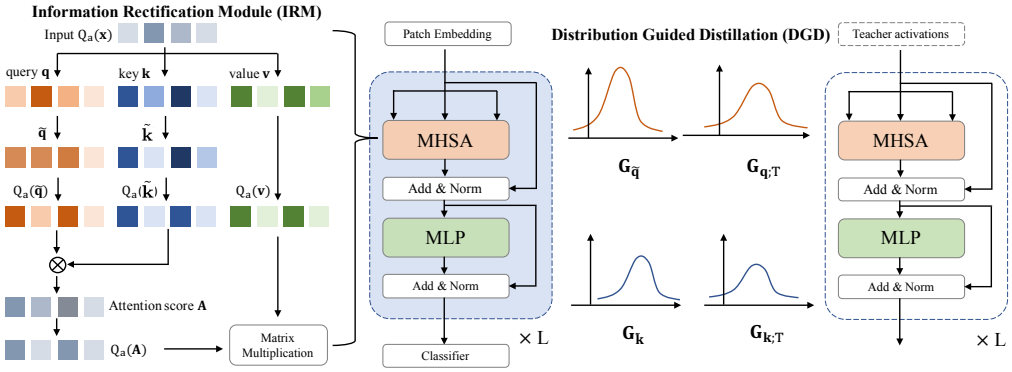


Figure 2: Overview of Q-ViT, applying Information Rectification Module (IRM) for maximizing representation information and Distribution Guided Distillation (DGD) for accurate optimization.

2 Related Work

Vision transformer. Motivated by the great success of Transformer in natural language processing, researchers are trying to apply Transformer architecture to Computer Vision tasks. Unlike mainstream CNN-based models, Transformer is capable of capturing long-distance visual relations by its self-attention module and provides the paradigm without image-specific inductive bias. ViT [4] views 16×16 image patches as token sequence and predicts classification via a unique class token, which shows promising results. Subsequently, many works, such as DeiT [24] and PVT [27] achieve further improvement on ViT, making it more efficient and applicable in downstream tasks. However, these high-performing vision transformers are attributed to the large number of parameters and high computational overhead, limiting their adoption. Therefore, innovating a smaller and faster vision transformer becomes a new trend. DynamicViT [22] presents a dynamic token sparsification framework to prune redundant tokens progressively and dynamically, achieving competitive complexity and accuracy trade-off. Evo-ViT [29] proposes a slow-fast updating mechanism that guarantees information flow and spatial structure, trimming down both the training and inference complexity. While the above works focus on efficient model designing, this paper boosts the compression and acceleration in the track of quantization.

Quantization. Quantizing neural networks (QNNs) often possess low-bit ($1\sim$ 4-bit) weights and activations to accelerate the model inference and save the memory usage. Specifically, ternary weights are introduced to reduce the quantization error in TWN [12]. DoReFa-Net [30] exploits convolution kernels with low bit-width parameters and gradients to accelerate both the training and inference. TTQ [31] uses two full-precision scaling coefficients to quantize the weights to ternary values. [32] presented a $2 \sim 4$-bit quantization scheme using a two-stage approach to alternately quantize the weights and activations, which provides an optimal trade-off among memory, efficiency, and performance. [10] parameterizes the quantization intervals and obtain their optimal values by directly minimizing the task loss of the network and also the accuracy degeneration with further bit-width reduction. [28] introduces transfer learning into network quantization to obtain an accurate low-precision model by utilizing the Kullback-Leibler (KL) divergence. [6] enables accurate approximation for tensor values that have bell-shaped distributions with long tails and finds the entire range by minimizing the quantization error. In our Q-ViT, we aim to implement an accurate, fully quantized vision transformer under the QAT paradigm.

3 Baseline of Fully Quantized ViT

First of all, we build a baseline to study the fully quantized ViT since it has never been proposed in previous works. A straightforward solution is to quantize the representations (weights and activations) in ViT architecture in the forward propagation and apply distillation to the optimization in the backward propagation.

Quantized ViT architecture. We briefly introduce the technology of neural network quantization. We first introduce a general asymmetric activation quantization and symmetric weight quantization

scheme as

$$
\begin{aligned}
Q_a(x) &= \lfloor \text{clip}\{(x-z)/\alpha_x, -Q_n^x, Q_p^x\} \rceil \quad & Q_w(\mathbf{w}) &= \lfloor \text{clip}\{\mathbf{w}/\alpha_{\mathbf{w}}, -Q_n^{\mathbf{w}}, Q_p^{\mathbf{w}}\} \rceil \\
\hat{x} &= Q_a(x) \times \alpha_x + z, & \hat{\mathbf{w}} &= Q_w(\mathbf{w}) \times \alpha_{\mathbf{w}}.
\end{aligned} \tag{1}
$$

Here, $\text{clip}\{y, r_1, r_2\}$ returns y with values below r_1 set as r_1 and values above r_2 set as r_2, and $\lfloor y \rceil$ rounds y to the nearest integer. With quantizing activations to signed a bits and weights to signed b bits, $Q_n^x = 2^{a-1}, Q_p^x = 2^{a-1} - 1$ and $Q_n^{\mathbf{w}} = 2^{b-1}, Q_p^{\mathbf{w}} = 2^{b-1} - 1$. In general, the forward and backward propagation of quantization function in quantized network is formulated as

$$
\begin{aligned}
&\text{Forward:} \quad \text{Q-Linear}(x) = \hat{x} \cdot \hat{\mathbf{w}} = \alpha_x \alpha_{\mathbf{w}}((Q_a(x) + z/\alpha_x) \otimes Q_w(\mathbf{w})), \\
&\text{Backward:} \quad \frac{\partial \mathcal{J}}{\partial x} = \frac{\partial \mathcal{J}}{\partial \hat{x}} \frac{\partial \hat{x}}{\partial x} = \begin{cases} \dfrac{\partial \mathcal{J}}{\partial \hat{x}} & \text{if } x \in [-Q_n^x, Q_p^x] \\ 0 & \text{otherwise} \end{cases}, \\
&\qquad\qquad \frac{\partial \mathcal{J}}{\partial \mathbf{w}} = \frac{\partial \mathcal{J}}{\partial x} \frac{\partial x}{\partial \hat{\mathbf{w}}} \frac{\partial \hat{\mathbf{w}}}{\partial \mathbf{w}} = \begin{cases} \dfrac{\partial \mathcal{J}}{\partial x} \dfrac{\partial x}{\partial \hat{\mathbf{w}}} & \text{if } \mathbf{w} \in [-Q_n^{\mathbf{w}}, Q_p^{\mathbf{w}}] \\ 0 & \text{otherwise} \end{cases},
\end{aligned} \tag{2}
$$

where $\mathcal{J}$ is loss function, $Q(\cdot)$ is applied in the forward propagation while the straight-through estimator (STE) [1] is used to retain the derivation of gradient in backward propagation. $\otimes$ denotes the matrix multiplication with efficient bit-wise operations.

The input images are first encoded as patches and passes through several transformer blocks. Such transformer block consists of two components: Multi-Head Self-Attention (MHSA) and Multi-Layer Perceptron (MLP). The computation of attention weight depends on the corresponding query $\mathbf{q}$, key $\mathbf{k}$ and value $\mathbf{v}$, and the quantized computation in one attention head is

$$
\mathbf{q} = \text{Q-Linear}_q(x), \mathbf{k} = \text{Q-Linear}_k(x), \mathbf{v} = \text{Q-Linear}_v(x), \tag{3}
$$

where $\text{Q-Linear}_q, \text{Q-Linear}_k, \text{Q-Linear}_v$ denote the three quantized linear layers for $\mathbf{q}, \mathbf{k}, \mathbf{v}$, respectively. Thus, the attention weight is formulated as

$$
\begin{aligned}
\mathbf{A} &= \frac{1}{\sqrt{d}}(Q_a(\mathbf{q}) \otimes Q_a(\mathbf{k})^\top), \\
Q_{\mathbf{A}} &= Q_a(\text{softmax}(\mathbf{A})).
\end{aligned} \tag{4}
$$

Training for Quantized ViT. Knowledge distillation is an essential supervision approach for training quantized neural networks, which bridges the performance gap between quantized models and full-precision counterparts. The usual practice is using distillation through attention as described in [24]

$$
\begin{aligned}
\mathcal{L}_{\text{dist}} &= \frac{1}{2}\mathcal{L}_{\text{CE}}(\psi(Z_q), y) + \frac{1}{2}\mathcal{L}_{\text{CE}}(\psi(Z_q), y_t), \\
y_t &= \arg\max_c Z_t(c).
\end{aligned} \tag{5}
$$

4 Proposed Q-ViT

With aforementioned quantizing and training pipeline, a fully quantized ViT baseline is built, which however has a large performance gap with full-precision counterparts. Our study shows that the baseline suffers an severe information distortion among quantized attention scores in the forward propagation and distillation direction misleading in the backward propagation.

4.1 Performance degeneration of Fully Quantized ViT Baseline

Intuitively, in the fully quantized ViT baseline, the information representation capability largely depends on the transformer-based architecture, such as the attention weight in MHSA module. However, the performance improvement brought by such architecture is severely limited by the quantized parameters, while the rounded and discrete quantization also significantly affect the optimization. The phenomenon identifies the bottleneck of the fully quantized ViT baseline comes from architecture and optimization for the forward and backward propagation, respectively.

Architecture bottleneck. We replace each module with the full-precision counterpart respectively and compare the accuracy drop as shown in Fig. 3. These ablation experiments are conducted the same as experiments in Tab. 2. We find that quantizing query, key, value and attention weight (*i.e.*, $\mathrm{softmax}(\mathbf{A})$ in Eq. (4) to 2 bits brings the most significant drop of accuracy among all parts of the ViT, up to 10.03%. While quantized MLP layers and quantized weights of linear layers in MHSA brings only 1.78% and 4.26% drop, respectively. And once query, key, value and attention weight are quantized, even keep all weights of linear layers in MHSA module full-precision, the performance drops (10.57%) are still significant. Thus, improving the attention structure is pivotal to solve the performance drop problem of quantized ViT.

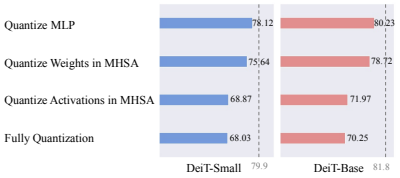


Figure 3: Analysis of bottleneck from architecture perspective. We report the accuracy of 2-bit quantized DeiT-S and DeiT-B on ImageNet dataset about replacing full-precision structure.

Optimization bottleneck. We calculate l2-norm distances between each attention weight among different blocks in DeiT-S architecture as shown in Fig. 4. The MHSA modules in full-precision ViT with different depth learn different representations from images. As mentioned in [21], lower ViT layers attend both locally and globally while higher ViT layers pay most attention to global representations. However, the fully quantized ViT (blue lines in Fig. 4) fails to learn accurate attention map distances. Thus, it requires a new design that could utilize the full-precision teacher's information better.

4.2 Information Rectification in Q-Attention

To address the information distortion of quantized representations in the forward propagation, we propose an efficient Q-Attention structure based on information theory, which statistically maximizes the entropy of representation and revives the attention mechanism in the fully quantized ViT. Since the representations with extremely compressed bit-width in fully quantized ViT have limited capabilities, the ideal quantized representation should preserve the given full-precision counterparts as much as possible, which means the mutual information between quantized and full-precision representations should be maximized as mentioned in [19].

We further show the statistical results that the distribution of query and key values in ViT architectures intending to follow Gaussian distributions under the distilling supervision, whose histograms are in bell-shape [19]. For example, in Fig. 1 and Fig. 5, we have shown the query and key distributions and their corresponding Probability Density Function (PDF) using the calculated mean and standard deviation for each MHSA layer. Therefore, the distributions of query and key in the MHSA modules of full-precision counterparts are formulated as

$$\mathbf{q} \sim \mathcal{N}(\mu(\mathbf{q}), \sigma(\mathbf{q})), \quad \mathbf{k} \sim \mathcal{N}(\mu(\mathbf{k}), \sigma(\mathbf{k})). \tag{6}$$

Since the weight and the activation with extremely compressed bit-width in fully quantized ViT have limited capabilities, the ideal quantization process should preserve the corresponding full-precision counterparts as much as possible, thus the mutual information between quantized and full-precision representations should be maximized [19]. As shown in [18], for Gaussian distribution, the quantizers with maximum output entropy (MOE) and minimum average error (MAE) are approximately the same within a multiplicative constant. Thus the process of minimizing the error between full-precision values and quantized values is equivalent to maximizing the information entropy of the quantized values. Thus, when the deterministic quantization function is applied to quantized ViT, such objective is equivalent to maximizing the information entropy $\mathcal{H}(Q_{\mathbf{x}})$ of quantized representation $Q_{\mathbf{x}}$ [18] in Eq. (4), which is defined as

$$\begin{aligned} \mathcal{H}(Q_a(\mathbf{x})) &= -\sum_{q_{\mathbf{x}}} p(q_{\mathbf{x}}) \log p(q_{\mathbf{x}}) = \frac{1}{2} \log 2\pi e \sigma_{\mathbf{x}}^2, \\ \max \mathcal{H}(Q_a(\mathbf{x})) &= \frac{n \ln 2}{2^n}, \quad \text{when } \ p(q_{\mathbf{x}}) = \frac{1}{2^n}, \end{aligned} \tag{7}$$

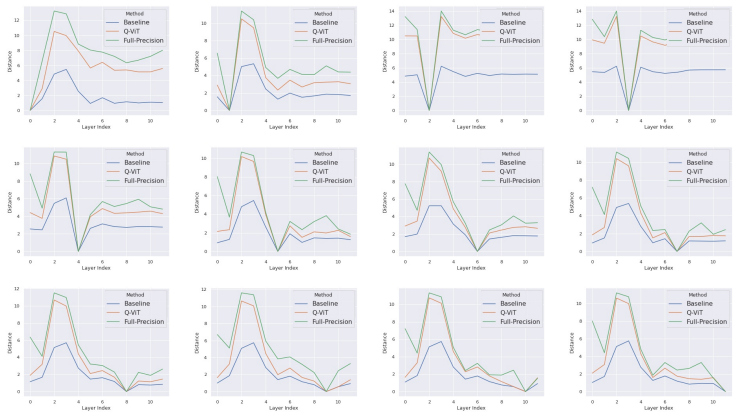


Figure 4: Attention-distance comparison for full-precision DeiT-Small (green lines) , fully quantized DeiT-Small baseline (blue lines), and Q-ViT (yellow lines) for same input. Q-ViT shows similar behavior with the full- precision model, while baseline suffers indistinguishable attention distance for information degradation.

where $q_{\mathbf{x}}$ is the random quantized variables in $Q_a(\mathbf{x})$ (which is $Q_a(\mathbf{q})$ or $Q_a(\mathbf{k})$ in different conditions) with probability mass function $p(\cdot)$. For better retaining the information contained in the MHSA modules from the full-precision counterparts, the information entropy in the quantization process should be maximized.

However, direct application of quantization function converting the values into finite fixed points brings irreversible disturbance to the distributions and the information entropy $\mathcal{H}(Q_a(\mathbf{q}))$ and $\mathcal{H}(Q_a(\mathbf{k}))$ degenerates to a much lower level than the full-precision counterparts. To mitigate the information degradation from the quantization process in the attention mechanism, a Information Rectification Module (IRM) is proposed for effectively maximizing the information entropy of quantized attention weights

$$Q_a(\tilde{\mathbf{q}}) = Q_a(\frac{\mathbf{q} - \mu(\mathbf{q}) + \beta_{\mathbf{q}}}{\gamma_{\mathbf{q}}\sqrt{\sigma^2(\mathbf{q}) + \epsilon_{\mathbf{q}}}}), \quad Q_a(\tilde{\mathbf{k}}) = Q_a(\frac{\mathbf{k} - \mu(\mathbf{k}) + \beta_{\mathbf{k}}}{\gamma_{\mathbf{k}}\sqrt{\sigma^2(\mathbf{k}) + \epsilon_{\mathbf{k}}}}), \tag{8}$$

where $\gamma_{\mathbf{q}}, \beta_{\mathbf{q}}$ and $\gamma_{\mathbf{k}}, \beta_{\mathbf{k}}$ are learnble parameters for modifying the distribution of $\tilde{\mathbf{q}}$, while $\epsilon_{\mathbf{q}}$ and $\epsilon_{\mathbf{k}}$ are constants preventing the denominator being 0. The learning rates of learnable $\gamma_{\mathbf{q}}, \beta_{\mathbf{q}}$ and $\gamma_{\mathbf{k}}, \beta_{\mathbf{k}}$ are same as the whole network. Thus after IRM, the information entropy $\mathcal{H}(Q_a(\tilde{\mathbf{q}}))$ and $\mathcal{H}(Q_a(\tilde{\mathbf{k}}))$ are formulated as

$$\mathcal{H}(Q(\tilde{\mathbf{q}})) = \frac{1}{2}\log 2\pi e[\gamma_{\mathbf{q}}^2(\sigma_{\mathbf{q}}^2 + \epsilon_{\mathbf{q}})], \quad \mathcal{H}(Q(\tilde{\mathbf{k}})) = \frac{1}{2}\log 2\pi e[\gamma_{\mathbf{k}}^2(\sigma_{\mathbf{k}}^2 + \epsilon_{\mathbf{k}})]. \tag{9}$$

Then to revive the attention mechanism to capture critic elements by information entropy maximization, the learnable parameters $\gamma_{\mathbf{q}}, \beta_{\mathbf{q}}$ and $\gamma_{\mathbf{k}}, \beta_{\mathbf{k}}$ reshape the distributions of the query and key values to achieve the state of information maximization. In a nutshell, in our IRM-Attention structure, the information entropy of quantized attention weight is maximized to alleviate its severe information distortion and revive the attention mechanism.

4.3 Distribution Guided Distillation through Attention

To address the attention distribution mismatch occurred in fully quantized ViT baseline in the backward propagation, we further propose a Distribution Guided Distillation (DGD) scheme with apposite distilled activations and the well-designed similarity matrices to effectively utilize knowledge from the teacher, which optimizes the fully quantized ViT more accurately.

As an optimization technique based on element-level comparison of activation, distillation allows the quantized ViT to mimic the full-precision teacher model about output logits. However, we find that the

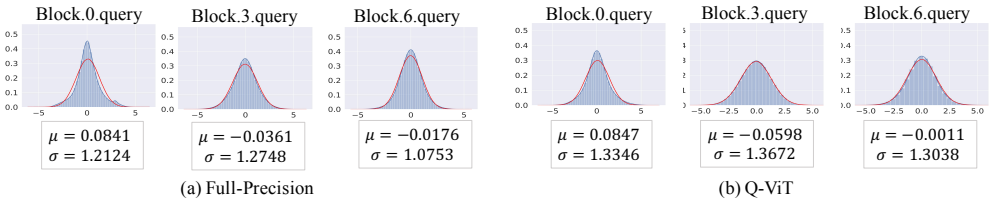


Figure 5: The histogram of query and key values $\mathbf{q}, \mathbf{k}$ (blue shadow) along with the PDF curve (red line) of Gaussian distribution $N(\mu, \sigma^2)$ [19], for 3 selected layers in DeiT-T and 4-bit Q-ViT. μ and σ^2 are the statistical mean and variance of the values.

distillation procedure used in previous ViT and fully quantized ViT baseline (Sec. 3) unable to deliver meticulous supervision to attention weights (shown in Fig. 4), leading to insufficient optimization. To solve the optimization insufficiency in the distillation of the fully quantized ViT, we propose the Distribution-Guided Distillation (DGD) method in Q-ViT. We first build patch-based similarity pattern matrices for distilling the upstream query and key instead of attention following [25], which is formulated as

$$\begin{aligned} \tilde{G}^l_{\mathbf{q}_h} &= \tilde{\mathbf{q}}^l_h \cdot (\tilde{\mathbf{q}}^l_h)^\top, \quad G^{(l)}_{\mathbf{q}_h} = \tilde{G}^l_{\mathbf{q}_h} / \|\tilde{G}^l_{\mathbf{q}_h}\|_2, \\ \tilde{G}^l_{\mathbf{k}_h} &= \tilde{\mathbf{k}}^l_h \cdot (\tilde{\mathbf{k}}^l_h)^\top, \quad G^{(l)}_{\mathbf{k}_h} = \tilde{G}^l_{\mathbf{k}_h} / \|\tilde{G}^l_{\mathbf{k}_h}\|_2, \end{aligned} \tag{10}$$

where $\|\cdot\|_2$ denotes ℓ_2 normalization, and l, h are layer index and head index. Previous work shows that matrices constructed in this way are regarded as the specific patterns reflecting the semantic comprehension of network [25]. And the patches encoded from input images containing high-level understanding of parts, objects, and scenes [7]. Thus such semantic-level distillation target provide guided and meticulous supervision to the quantized ViT. The corresponding $\tilde{G}^l_{\mathbf{q}_h;T}$ and $\tilde{G}^l_{\mathbf{k}_h;T}$ are constructed in the same way by the teacher's activation. Thus combining the original distillation loss in Eq. (5), the final distillation loss is then formulated as

$$\begin{aligned} \mathcal{L}_{\text{DGD}} &= \sum_{l\in[1,L]} \sum_{h\in[1,H]} \|\tilde{G}^l_{\mathbf{q}_h;T} - \tilde{G}^l_{\mathbf{q}_h}\|_2 + \|\tilde{G}^l_{\mathbf{k}_h;T} - \tilde{G}^l_{\mathbf{k}_h}\|_2, \\ \mathcal{L}_{\text{distillation}} &= \mathcal{L}_{\text{dist}} + \mathcal{L}_{\text{DGD}}, \end{aligned} \tag{11}$$

where L and H denote the number of ViT layers and heads. With the proposed Distribution Guided Distillation, the Q-ViT retains the distribution over query and key from the full-precision counterparts (as shown in Fig. 5).

Our DGD scheme first provides the distribution-aware optimization direction together with processing appropriate distilled parameters and then constructs similarity matrices to eliminate scale differences and numerical instability, thereby improves fully quantized ViT by accurate optimization.

5 Experiments

In this section, we evaluate the performance of the proposed Q-ViT model for image classification task using popular DeiT [24] and Swin [14] backbones. To the best of our knowledge, there is no publicly available source code on quantization-aware training of vision transformer at this point, so we implement the baseline and LSQ [5] methods by ourselves.

5.1 Datasets and Implementation Details

Datasets. The experiments are carried out on the ILSVRC12 ImageNet classification dataset [11]. The ImageNet dataset is more challenging due to its large scale and greater diversity. There are 1000 classes and 1.2 million training images, and 50k validation images in it. In our experiments, we use the classic data augmentation method described in [24].

Experimental settings. In our experiments, we initialize the weights of quantized model with the corresponding pretrained full-precision model. The quantized model is trained for 300 epochs with batch-size 512 and the base learning rate $2e-4$. We do not use warm-up scheme. For all the

experiments, we apply Adam optimizer with weight decay set as 0. Other training settings follow DeiT [24] or Swin Transformer [14]. Note that we use 8-bit for the patch embedding (first) layer and the classification (last) layer following [5].

Bakcbone. We evaluate our quantization method on two popular vision transformer implementation: DeiT [24] and Swin Transformer [14]. The DeiT-S, DeiT-B, Swin-T and Swin-S are adopted as the backbone models, whose Top-1 accuracy on ImageNet dataset are 79.9%, 81.8%, 81.2%, and 83.2% respectively. For a fair comparison, we utilize the official implementation of DeiT and Swin Transformer.

5.2 Ablation Study

We give quantitative results of the proposed IRM and DGD in Tab. 1 As shown in Tab. 1, the fully quantized ViT baseline suffers a severe performance drop on classification task (0.2%, 2.1% and 11.7% with 2/3/4-bit, respectively). IRM and DGD improve the performance when used alone, and the two techniques further boost the performance considerably when combined together. For example, the IRM improve the 2-bit Baseline by 1.7% and the DGD achieves 2.3% performance improvement. While combining the IRM and DGD together, the performance improvement achieves 3.8%.

Table 1: Evaluating the components of Q-ViT based on ViT-S backbone.

Method	#Bits	Top-1	#Bits	Top-1	#Bits	Top-1
Full-precision	32-32	79.9	-	-	-	-
Baseline	4-4	79.7	3-3	77.8	2-2	68.2
+IRM	4-4	80.2	3-3	78.2	2-2	69.9
+DGD	4-4	80.4	3-3	78.5	2-2	70.5
+IRM+DGD (Q-ViT)	4-4	**80.9**	3-3	**79.0**	2-2	**72.0**

To conclude, the two techniques can promote each other to improve Q-ViT and close the performance gap between fully quantized ViT and full-precision counterpart. More detailed ablation experiments and visualization reults can be found in Supplementary.

5.3 Main Results

The experimental results are shown in Tab. 2. We compare our method with 2/3/4-bit baseline and LSQ [5] based on the same frameworks for the task of image classification with the ImageNet dataset. We also report the classification performance of the 8-bit post-training quantization networks percentile VT-PTQ [17]. We firstly evaluate the proposed method on DeiT-S and DeiT-B models.

For DeiT-S backbone, compared with 8-bit VT-PTQ method, our 4bit Q-ViT achieves a much larger compression ratio than 8-bit VT-PTQ, but with significant performance improvement (78.1% *vs.* 80.9%). And it is worth noting that the proposed 2-bit model significantly compresses the DeiT-S by 21.5× on FLOPs. The proposed method boosts the performance of 2/3/4-bit Baseline by 3.9%, 1.5% and 1.2% with the same architecture and bit-width, which is significant on the ImageNet dataset. For larger DeiT-B, as shown in Tab. 2, the performance of the proposed method outperforms the 2/3/4-bit Baseline by 3.8%, 1.7% and 1.9%, a large margin. Also note that the proposed 2/3/4-bit model significantly compresses the DeiT-B by 21×, 12× and 7.6× on FLOPs. Compared with 8-bit post-training quantization methods, our method achieves significantly higher compression rate, and the performance improvement is significant.

Also, our method generates convincing results on Swin-transformers. As shown in Tab. 2, the performance of the proposed method with Swin-T outperforms the 2/3/4-bit Baseline method by 4.1% , 2.1% and 2.0%, a large margin. Compared with 8-bit post-training quantization methods, our method achieves significantly higher compression rate, and comparable performance. Note that our 4-bit Q-ViT surpasses the full-precision by 1.3% counterpart using Swin-T, which demonstrates the significance of our Q-ViT. For larger Swin-S, the performance of the proposed method outperforms the 2/3/4-bit Baseline by 4.3%, 1.8% and 1.5%. Also note that our 4-bit Q-ViT surpasses the full-

Table 2: Quantization results on ImageNet dataset. "#Bits" (W-A) is the bit width for weights and activation.

Network	Method	#Bits	$\text{Size}_{(\text{MB})}$	$\text{FLOPs}_{(\text{G})}$	Top-1	Top-5
DeiT-S	Full-precision	32-32	88.2	4.3	79.9	95.0
	VT-PTQ	8_{MP}-8_{MP}	22.2	-	78.1	-
	LSQ	4-4	11.4	0.7	79.6	94.6
	Baseline	4-4	11.4	0.7	79.7	94.5
	Q-ViT	4-4	11.4	0.7	**80.9**	**94.9**
	LSQ	3-3	8.7	0.4	77.3	93.0
	Baseline	3-3	8.7	0.4	77.5	93.3
	Q-ViT	3-3	8.7	0.4	**79.0**	**94.2**
	LSQ	2-2	6.0	0.2	68.0	86.4
	Baseline	2-2	6.0	0.2	68.2	86.5
	Q-ViT	2-2	6.0	0.2	**72.1**	**90.3**
DeiT-B	Full-precision	32-32	346.2	16.8	81.8	95.6
	VT-PTQ	8_{MP}-8_{MP}	86.8	-	81.3	-
	LSQ	4-4	44.1	2.2	80.9	95.1
	Baseline	4-4	44.1	2.2	81.1	95.3
	Q-ViT	4-4	44.1	2.2	**83.0**	**96.1**
	LSQ	3-3	33.4	1.4	79.0	94.5
	Baseline	3-3	33.4	1.4	79.3	94.9
	Q-ViT	3-3	33.4	1.4	**81.0**	**95.1**
	LSQ	2-2	22.7	0.8	70.3	88.6
	Baseline	2-2	22.7	0.8	70.4	88.8
	Q-ViT	2-2	22.7	0.8	**74.2**	**92.2**
Swin-T	Full-precision	32-32	114.2	4.5	81.2	95.5
	LSQ	4-4	14.6	0.6	80.2	95.2
	Baseline	4-4	14.6	0.6	80.5	95.4
	Q-ViT	4-4	14.6	0.6	**82.5**	**97.3**
	LSQ	3-3	11.2	0.3	79.7	94.9
	Baseline	3-3	11.2	0.3	79.8	95.1
	Q-ViT	3-3	11.2	0.3	**80.9**	**96.1**
	LSQ	2-2	7.7	0.2	70.4	88.8
	Baseline	2-2	7.7	0.2	70.6	89.0
	Q-ViT	2-2	7.7	0.2	**74.7**	**92.5**
Swin-S	Full-precision	32-32	199.8	8.7	83.2	96.2
	LSQ	4-4	7.0	1.1	82.5	97.1
	Baseline	4-4	7.0	1.1	82.9	97.3
	Q-ViT	4-4	7.0	1.1	**84.4**	**98.3**
	LSQ	3-3	5.5	0.6	80.6	95.7
	Baseline	3-3	5.5.	0.6	80.9	95.9
	Q-ViT	3-3	5.5	0.6	**82.7**	**97.5**
	LSQ	2-2	3.9	0.3	72.4	90.2
	Baseline	2-2	3.9	0.3	72.7	90.6
	Q-ViT	2-2	3.9.	0.3	**76.9**	**94.9**

precision by 1.1% counterpart using Swin-S and significantly compresses the Swin-S by 7.9$\times$, which demonstrates the effectiveness and efficiency of our Q-ViT.

6 Conclusion

In this paper, we introduce Q-ViT to improve the fully quantized ViTs with high compression ratio and competitive performance. We first build a theoretical framework of fully quantized ViT and analysis the bottlenecks of the fully quantized ViT baseline. Then we introduce Information Rectification Module and Distribution Guided Distillation to Q-ViT for performance improvement. Our proposed Q-ViTs achieve comparable performance with full-precision counterparts with ultra-low bit weights and activations. Our work gives an insightful analysis and effective solution about the crucial issues in ViT full quantization, which blazes a promising path for the extreme compression of ViT.

References

[1] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proc. of ECCV*, pages 213–229, 2020.

[3] Misha Denil, Babak Shakibi, Laurent Dinh, Marc'Aurelio Ranzato, and Nando De Freitas. Predicting parameters in deep learning. In *Proc. of NeurIPS*, pages 2148–2156, 2013.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[5] Steven K Esser, Jeffrey L McKinstry, Deepika Bablani, Rathinakumar Appuswamy, and Dharmendra S Modha. Learned step size quantization. *arXiv preprint arXiv:1902.08153*, 2019.

[6] Jun Fang, Ali Shafiee, Hamzah Abdel-Aziz, David Thorsley, Georgios Georgiadis, and Joseph H Hassoun. Post-training piecewise linear quantization for deep neural networks. In *Proc. of ECCV*, pages 69–86, 2020.

[7] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.

[8] Yang He, Guoliang Kang, Xuanyi Dong, Yanwei Fu, and Yi Yang. Soft filter pruning for accelerating deep convolutional neural networks. In *Proc. of IJCAI*, pages 2234–2240, 2018.

[9] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[10] Sangil Jung, Changyong Son, Seohyung Lee, Jinwoo Son, Jae-Joon Han, Youngjun Kwak, Sung Ju Hwang, and Changkyu Choi. Learning to quantize deep networks by optimizing quantization intervals with task loss. In *Proc. of CVPR*, pages 4350–4359, 2019.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. of NeurIPS*, pages 1097–1105, 2012.

[12] Fengfu Li, Bo Zhang, and Bin Liu. Ternary weight networks. *arXiv preprint arXiv:1605.04711*, 2016.

[13] Yang Lin, Tianyu Zhang, Peiqin Sun, Zheng Li, and Shuchang Zhou. Fq-vit: Fully quantized vision transformer without retraining. *arXiv preprint arXiv:2111.13824*, 2021.

[14] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proc. of ICCV*, pages 10012–10022, 2021.

[15] Zechun Liu, Zhiqiang Shen, Marios Savvides, and Kwang-Ting Cheng. Reactnet: Towards precise binary neural network with generalized activation functions. In *Proc. of ECCV*, pages 143–159, 2020.

[16] Zechun Liu, Baoyuan Wu, Wenhan Luo, Xin Yang, Wei Liu, and Kwang-Ting Cheng. Bi-real net: Enhancing the performance of 1-bit cnns with improved representational capability and advanced training algorithm. In *Proc. of ECCV*, 2018.

[17] Zhenhua Liu, Yunhe Wang, Kai Han, Wei Zhang, Siwei Ma, and Wen Gao. Post-training quantization for vision transformer. *Proc. of NeurIPS*, 2021.

[18] D Messerschmitt. Quantizing for maximum output entropy (corresp.). *IEEE Transactions on Information Theory*, 17(5):612–612, 1971.

[19] Haotong Qin, Yifu Ding, Mingyuan Zhang, Qinghua Yan, Aishan Liu, Qingqing Dang, Ziwei Liu, and Xianglong Liu. Bibert: Accurate fully binarized bert. In *Proc. of ICLR*, pages 1–24, 2022.

[20] Haotong Qin, Ruihao Gong, Xianglong Liu, Mingzhu Shen, Ziran Wei, Fengwei Yu, and Jingkuan Song. Forward and backward information retention for accurate binary neural networks. In *Proc. of CVPR*, pages 2250–2259, 2020.

[21] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Proc. of NeurIPS*, pages 1–13, 2021.

[22] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Proc. of NeurIPS*, pages 1–14, 2021.

[23] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In *Proc. of ICLR*, pages 1–13, 2015.

[24] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proc. of ICML*, pages 10347–10357, 2021.

[25] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proc. of ICCV*, pages 1365–1374, 2019.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Proc. of NeurIPS*, 30, 2017.

[27] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proc. of ICCV*, pages 568–578, 2021.

[28] Zheng Xie, Zhiquan Wen, Jing Liu, Zhiqiang Liu, Xixian Wu, and Mingkui Tan. Deep transferring quantization. In *Proc. of ECCV*, pages 625–642, 2020.

[29] Yifan Xu, Zhijie Zhang, Mengdan Zhang, Kekai Sheng, Ke Li, Weiming Dong, Liqing Zhang, Changsheng Xu, and Xing Sun. Evo-vit: Slow-fast token evolution for dynamic vision transformer. *arXiv preprint arXiv:2108.01390*, 2021.

[30] Shuchang Zhou, Yuxin Wu, Zekun Ni, Xinyu Zhou, He Wen, and Yuheng Zou. Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. *arXiv preprint arXiv:1606.06160*, 2016.

[31] Chenzhuo Zhu, Song Han, Huizi Mao, and William J Dally. Trained ternary quantization. In *Proc. of ICLR*, pages 1–10, 2017.

[32] Bohan Zhuang, Chunhua Shen, Mingkui Tan, Lingqiao Liu, and Ian Reid. Towards effective low-bitwidth convolutional neural networks. In *Proc. of CVPR*, pages 7920–7928, 2018.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section 5.

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 4.
   (b) Did you describe the limitations of your work? [Yes] Q-ViT is effective and efficient as shown in our extensive experiments, but the extension to 1 bit is still unexplored. We believe that the knowledge distillation, data augmentation can be used to to further enhance the performance of 1-bit Q-ViT, which will be further explored in our future work.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] This work does not have a direct negative social impact. However, we should be aware of the power of neural networks and prevent them from being used for harmful applications.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowd sourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]