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

Yanjing Li†, Sheng Xu†, Baochang Zhang, Xianbin Cao*, Peng Gao, Guodong Guo

{yanjingli, shengxu, bczhang, xbcao}@buaa.edu.cn

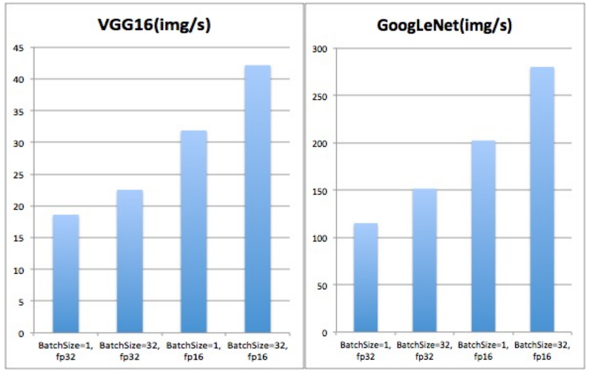
NeurIPS 2022

Motivation

- **Constraint of ViT Applications: Huge FLOPs**

Model	FLOPs	Memory Usage
ViT [1]-H	162GB	2528MB
DeiT[2]-B	16.8GB	346.2MB
Swin[3]-S	8.7GB	199.8MB

- **Deploying NNs on NVIDIA Jetson TX2 [4] :**



non real-time computation

*MB=1024^2bit, GB=1024^3bit

[1] Alexey Dosovitskiy, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929, 2020
[2] Hugo Touvron, Matthieu Cord,et al. Training data-efficient image transformers & distillation through attention. In Proc. of ICML, 2020
[3] Ze Liu, Yutong Lin, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In Proc. Of ICCV, 2020
[4] *https://www.nvidia.cn/autonomous-machines/embedded-systems/jetson-tx2/*

Baseline of Quantized ViT

1. Quantized ViT scheme

Symmetric weight quantization:

$$Q_w(\mathbf{w}) = \lfloor \mathrm{clip}\{\frac{\mathbf{w}}{\alpha_{\mathbf{w}}}, -Q_n^{\mathbf{w}}, -Q_p^{\mathbf{w}}\} \rceil$$

$$\hat{\mathbf{w}} = Q_w(\mathbf{w}) \times \alpha_{\mathbf{w}}$$

Asymmetric activation quantization:

$$Q_a(x) = \lfloor \mathrm{clip}\{\frac{x - z}{\alpha_x}, -Q_n^x, -Q_p^x\} \rceil$$

$$\hat{x} = Q_a(x) \times \alpha_x + z$$

Baseline of Quantized ViT

2. Quantized MHSA

MLP layer quantization:

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

Attention weight quantization:

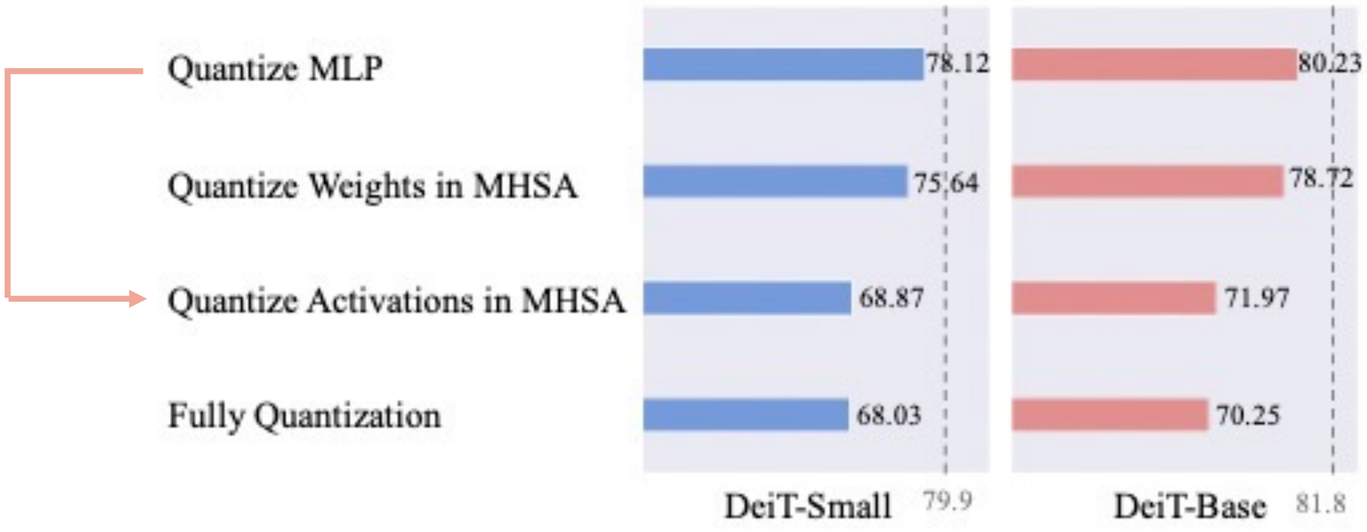
$$\mathbf{A} = \frac{1}{\sqrt{d}}\left(Q_a(\mathbf{q}) \otimes Q_a(\mathbf{k})^T\right)$$

$$Q_{\mathbf{A}} = Q_a(\text{softmax}(\mathbf{A}))$$

Baseline of Quantized ViT

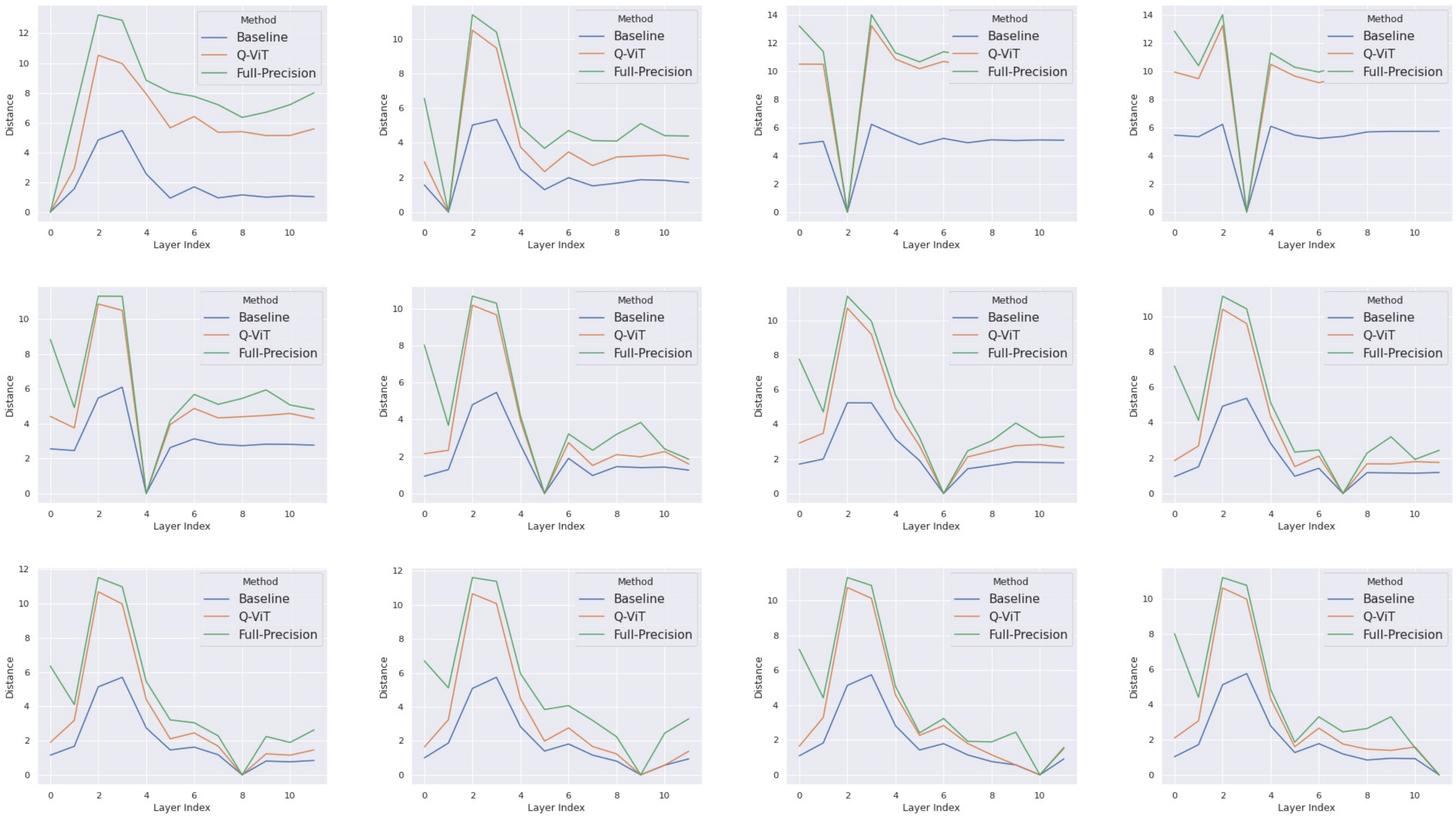
3. Quantized ViT Architecture Bottleneck

Quantizing query, key, value and attention weight brings the most significant drop

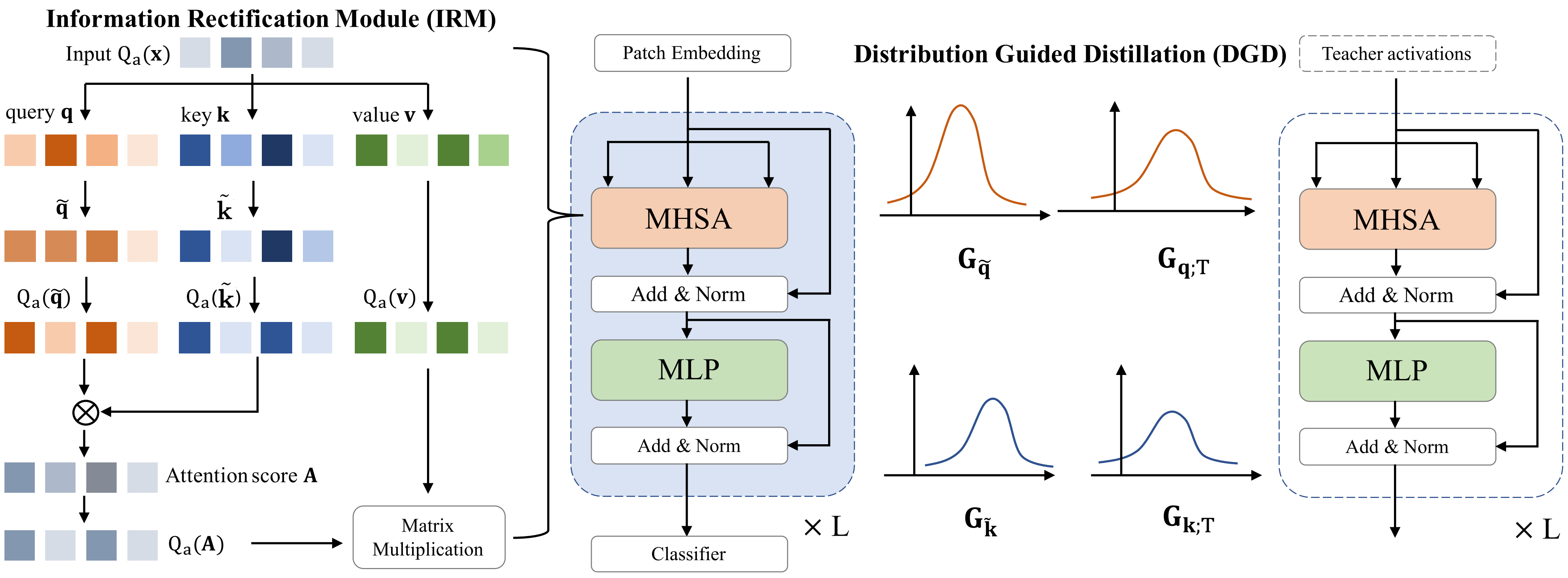


Baseline of Quantized ViT

4. Quantized ViT Optimization Bottleneck



Framework and Proposed Q-ViT



Framework and Proposed Q-ViT

1. Information Rectification Module -> Solving the Architecture Bottleneck

Information rectification

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

Information entropy maximization

$$\mathcal{H}(Q_a(\tilde{\mathbf{q}})) = \frac{1}{2}\log 2\pi e\left[\gamma_{\mathbf{q}}^2(\sigma^2(\mathbf{q}) + \epsilon_{\mathbf{q}})\right], \mathcal{H}(Q_a(\tilde{\mathbf{k}})) = \frac{1}{2}\log 2\pi e\left[\gamma_{\mathbf{k}}^2(\sigma^2(\mathbf{k}) + \epsilon_{\mathbf{k}})\right]$$

Framework and Proposed Q-ViT

2. Distributed Guided Distillation -> Solving the Optimization Bottleneck

Patch-based similarity in query and key

$$\tilde{G}_{\mathbf{q}_h}^{l} = \tilde{\mathbf{q}}_h^{l} \cdot \left(\tilde{\mathbf{q}}_h^{l}\right)^{\top}, G_{\mathbf{q}_h}^{(l)} = \frac{\tilde{G}_{\mathbf{q}_h}^{l}}{\left\|\tilde{G}_{\mathbf{q}_h}^{l}\right\|_2}$$

$$\tilde{G}_{\mathbf{k}_h}^{l} = \tilde{\mathbf{k}}_h^{l} \cdot \left(\tilde{\mathbf{k}}_h^{l}\right)^{\top}, G_{\mathbf{k}_h}^{(l)} = \frac{\tilde{G}_{\mathbf{k}_h}^{l}}{\left\|\tilde{G}_{\mathbf{k}_h}^{l}\right\|_2}$$

Final distillation loss

$$\mathcal{L}_{\mathrm{DGD}} = \sum_{l\in[1,L]} \sum_{h\in[1,H]} \left\|G_{\mathbf{q}_h;T}^{(l)} - G_{\mathbf{q}_h}^{(l)}\right\|_2 + \left\|G_{\mathbf{k}_h;T}^{(l)} - G_{\mathbf{k}_h}^{(l)}\right\|_2$$

Experiments and Results

Ablation Study

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**

- The fully quantized ViT **baseline** suffers a **severe performance drop** on classification task (11.7%, 2.1% and 0.2% with 2/3/4-bit, respectively).
- 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%**.

Experiments and Results

Main Results

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**

For DeiT-S:

- 4bit Q-ViT **surpasses full-precision** DeiT-S (**80.9% vs. 79.9%**).
- 2-bit model significantly compresses the DeiT-S by **21.5x** on **FLOPs**.

For larger DeiT-B:

- Q-ViT outperforms the 2/3/4-bit Baseline by **3.8%, 1.7% and 1.9%** , a large margin.
- 2/3/4-bit Q-ViT significantly compresses the DeiT-B by **21x, 12x and 7.6x** on **FLOPs**.

Experiments and Results

Main Results

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

Network	Method	#Bits	Size$_{(MB)}$	FLOPs$_{(G)}$	Top-1	Top-5
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**

For Swin-T:

- Q-Swin-T outperforms the 2/3/4-bit Baseline method by **4.1% , 2.1% and 2.0%** , a large margin.
- Our 4-bit Q-ViT **surpasses the full-precision** Swin-T by **1.3%**.

For larger Swin-S:

- Our method outperforms the 2/3/4-bit Baseline by **4.3%, 1.8% and 1.5%** .
- 4-bit Q-ViT surpasses the full-precision by **1.1%** counterpart using Swin-S and significantly compresses the Swin-S by **7.9x** on **FLOPs**.

Experiments and Results

Quantitative Results

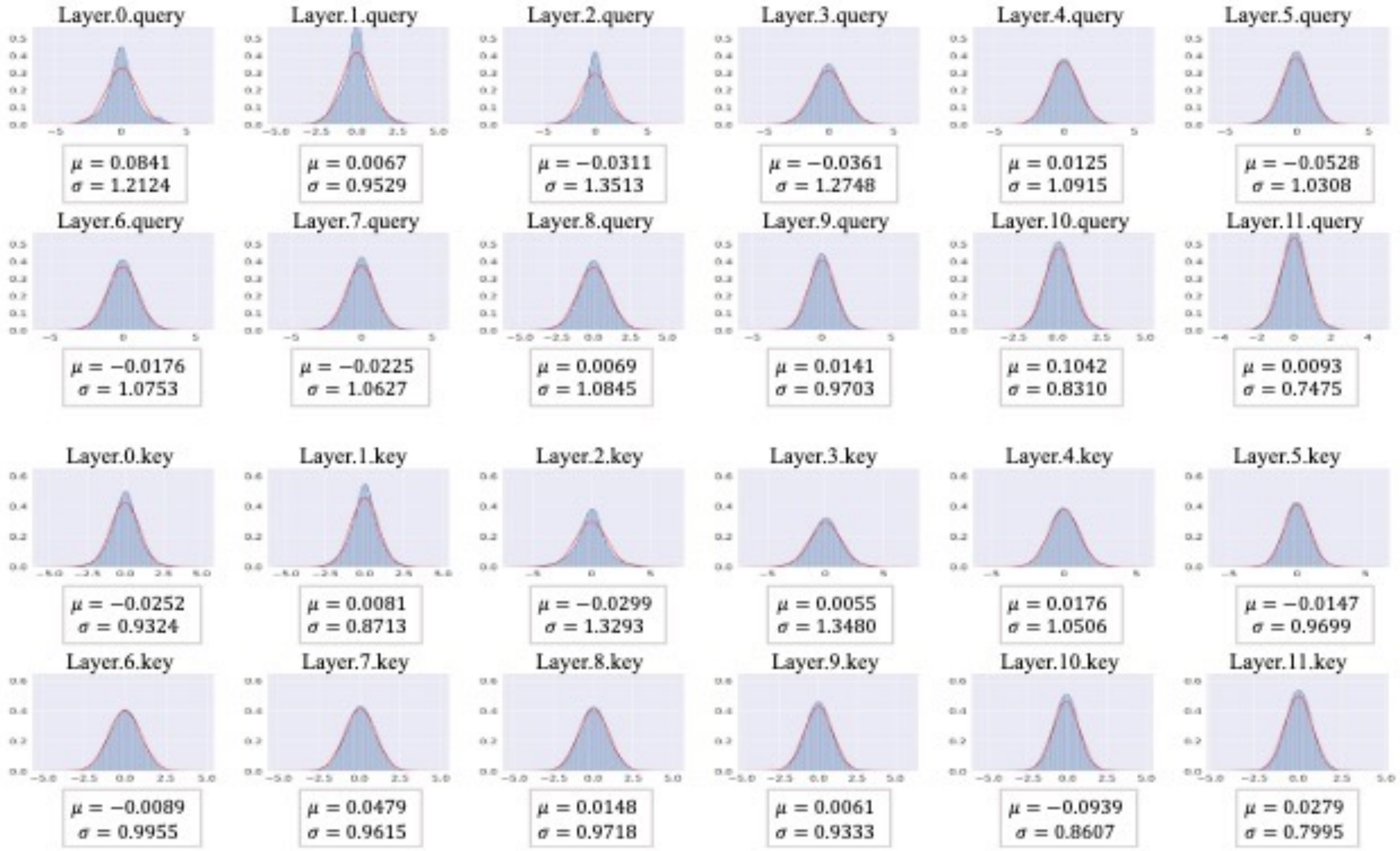


Figure 1: The histogram of query and key values **q**, **k** (blue shadow) along with the PDF curve (red line) of Gaussian distribution $N(\mu, \sigma^2)$ [2], for all 12 layers in full-preciison DeiT-T. μ and σ^2 are the statistical mean and variance of the values.

Experiments and Results

Quantitative Results

With the proposed IRM and DGD, the Q-ViT **retains the distribution over query and key** from the full-precision counterpart.

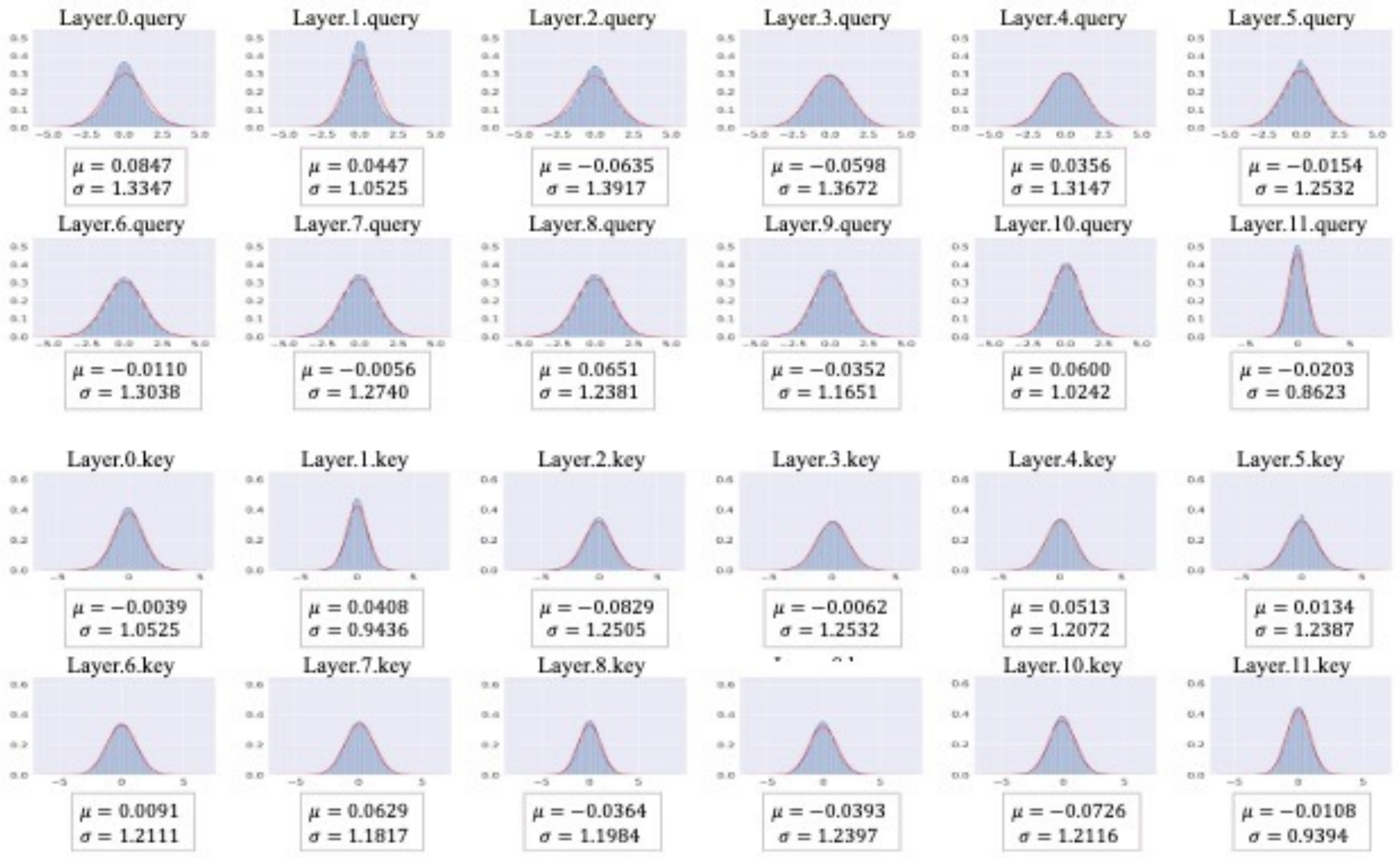


Figure 3: The histogram of query and key values **q**, **k** (blue shadow) along with the PDF curve (red line) of Gaussian distribution $N(\mu, \sigma^2)$ [2], for all 12 layers in Q-ViT. μ and σ^2 are the statistical mean and variance of the values.

Conclusion

- 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.
- We then introduce Information Rectification Module and Distribution Guided Distillation to Q-ViT for performance improvement.
- 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.
- Our proposed Q-ViTs achieve comparable performance with full-precision counterparts with ultra-low bit weights and activations.

Thank you for listening