TUE-AM-366

Q-DETR: An Efficient Low-Bit Quantized Detection Transformer

Sheng Xu[1,†], Yanjing Li[1,†], Mingbao Lin[3], Peng Gao[4], Guodong Guo[5], Jinhu Lü[1,2], Baochang Zhang[1,2],

[1]Beihang University, [2]Zhongguancun Laboratory, [3]Tencent, [4]Shanghai AI Laboratory, [5]UNIUBI Research, Universal Ubiquitous Co.

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

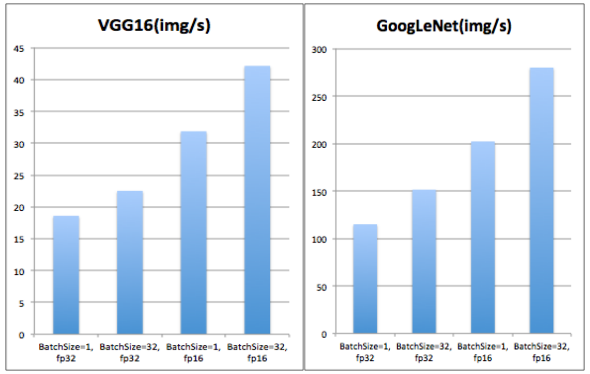
CVPR2023 Highlight

Background

- **Constraint of ViT & DETR 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=1024²bit, GB=1024³bit

[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 DETR

1. Quantized DETR scheme

Symmetric weight quantization:

$$\mathbf{w}_q = \lfloor \mathrm{clip}\{\frac{\mathbf{w}}{\alpha_{\mathbf{w}}}, Q_n^{\mathbf{w}}, Q_p^{\mathbf{w}}\} \rceil,$$

$$Q_w(x) = \alpha_{\mathbf{w}} \circ \mathbf{w}_q,$$

Asymmetric activation quantization:

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

$$Q_a(x) = \alpha_x \circ \boldsymbol{x}_q + z,$$

Baseline of Quantized DETR

2. Quantized MHA

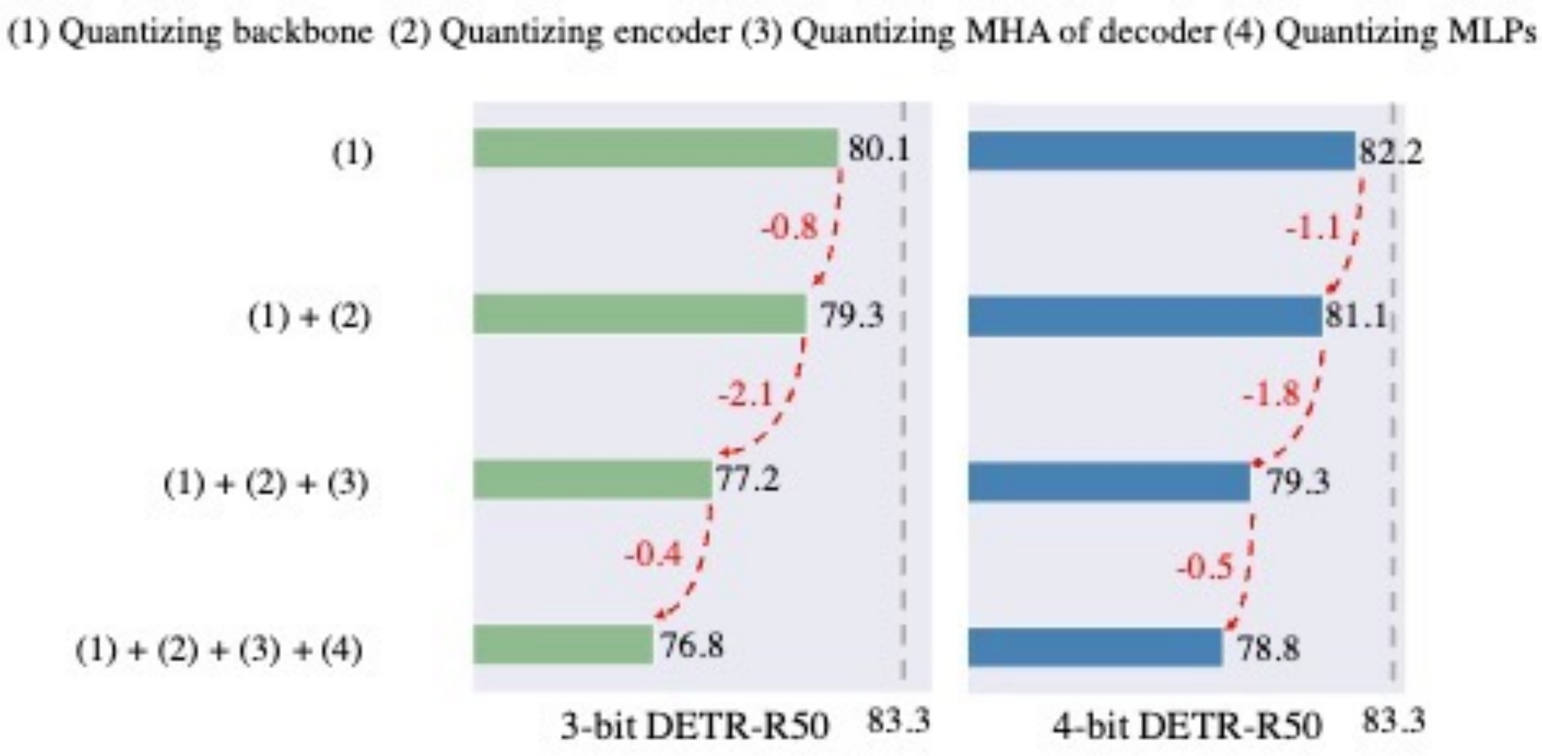
$$\text{Q-FC}(\boldsymbol{x}) = Q_a(\boldsymbol{x}) \cdot Q_w(\mathbf{w}) = \alpha_x \alpha_{\mathbf{w}} \circ (\boldsymbol{x}_q \odot \mathbf{w}_q + z/\alpha_x \circ \mathbf{w}_q),$$

$$\begin{aligned}
\mathbf{q} &= \text{Q-FC}(\mathbf{O}), \quad \mathbf{k}, \mathbf{v} = \text{Q-FC}(\mathbf{E}) \\
\mathbf{A}_i &= \text{softmax}(Q_a(\mathbf{q})_i \cdot Q_a(\mathbf{k})_i^\top / \sqrt{d}), \\
\mathbf{D}_i &= Q_a(\mathbf{A})_i \cdot Q_a(\mathbf{v})_i,
\end{aligned}$$

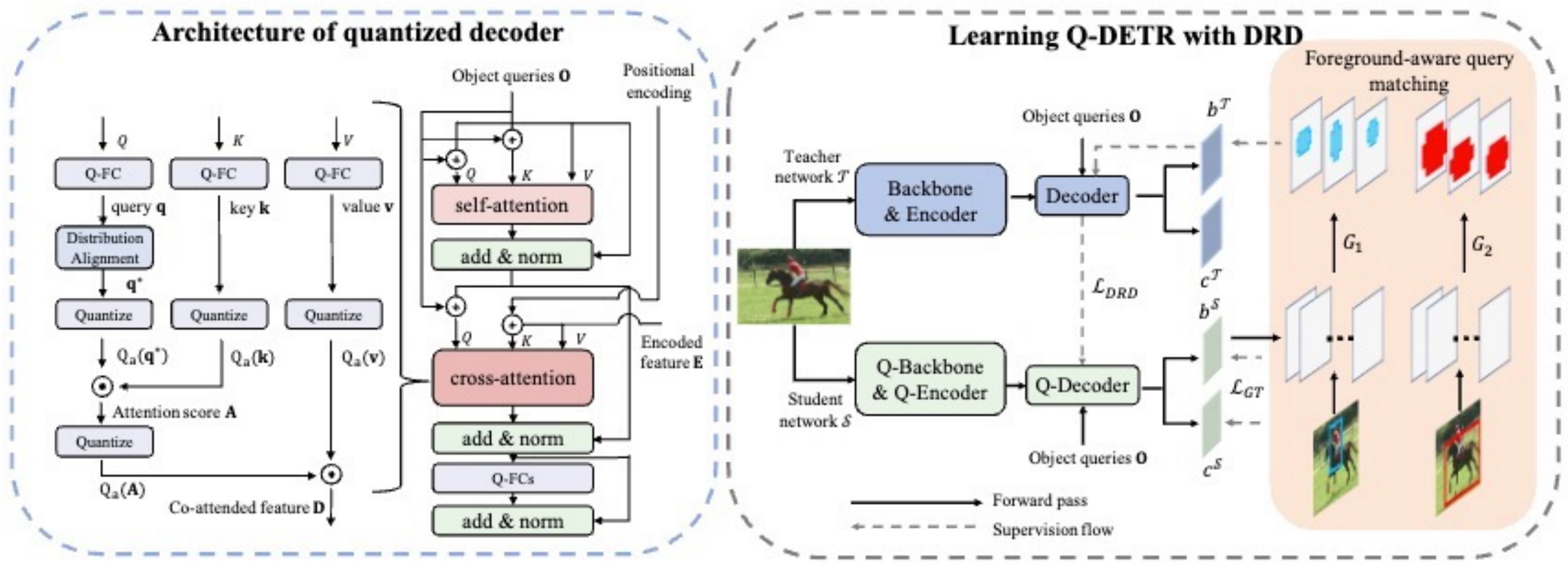
Baseline of Quantized DETR

3. Challenge Analysis

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



Framework and Proposed Q-DETR



Framework and Proposed Q-DETR

1. Information Bottleneck of Q-DETR

$$\min_{\theta^{\mathcal{S}}} I(X;\mathbf{E}^{\mathcal{S}}) - \beta I(\mathbf{E}^{\mathcal{S}}, \mathbf{q}^{\mathcal{S}}; \boldsymbol{y}^{GT}) - \gamma I(\mathbf{q}^{\mathcal{S}}; \mathbf{q}^{\mathcal{T}})$$

$$I(\mathbf{q}^{\mathcal{S}}; \mathbf{q}^{\mathcal{T}}) = H(\mathbf{q}^{\mathcal{S}}) - H(\mathbf{q}^{\mathcal{S}} | \mathbf{q}^{\mathcal{T}})$$

$$\min_{\theta} H(\mathbf{q}^{\mathcal{S}^*} | \mathbf{q}^{\mathcal{T}}),$$

$$\text{s. t.} \quad \mathbf{q}^{\mathcal{S}^*} = \arg\max_{\mathbf{q}^{\mathcal{S}}} H(\mathbf{q}^{\mathcal{S}})$$

Framework and Proposed Q-DETR

2. Distribution Rectification Distillation

Inner-level optimization:

$$H(\mathbf{q}^{\mathcal{S}}) = -\int_{\mathbf{q}_i^{\mathcal{S}} \in \mathbf{q}^{\mathcal{S}}} p(\mathbf{q}_i^{\mathcal{S}}) \log p(\mathbf{q}_i^{\mathcal{S}})$$

$$\begin{aligned} H(\mathbf{q}^{\mathcal{S}}) &= -\mathbb{E}[\log \mathcal{N}(\mu(\mathbf{q}^{\mathcal{S}}), \sigma(\mathbf{q}^{\mathcal{S}}))] \\ &= -\mathbb{E}[\log[(2\pi\sigma(\mathbf{q}^{\mathcal{S}})^2)^{\frac{1}{2}} \exp(-\frac{(\mathbf{q}_i^{\mathcal{S}} - \mu(\mathbf{q}^{\mathcal{S}}))^2}{2\sigma(\mathbf{q}^{\mathcal{S}})^2})]] \\ &= \frac{1}{2} \log 2\pi\sigma(\mathbf{q}^{\mathcal{S}})^2. \end{aligned}$$

Framework and Proposed Q-DETR

2. Distribution Rectification Distillation

Upper-level optimization:

$$G_i = \max_{1 \le j \le N} \mathrm{GIoU}(b_i^{GT}, b_j^{\mathcal{S}}),$$

$$b_j^{\mathcal{S}} = \begin{cases} b_j^{\mathcal{S}}, & \mathrm{GIoU}(b_i^{GT}, b_j^{\mathcal{S}}) > \tau G_i, \ \forall\, i \\ \varnothing, & \text{otherwise}, \end{cases}$$

$$\tilde{c}_j^{\mathcal{T}}, \tilde{b}_j^{\mathcal{T}} = \arg\max_{\tilde{c}_k^{\mathcal{T}}, \tilde{b}_k^{\mathcal{T}}} \sum_{k=1}^{N} \mu_1 \mathrm{GIoU}(\tilde{b}_j^{\mathcal{S}}, b_k^{\mathcal{T}}) - \mu_2 \|\tilde{b}_j^{\mathcal{S}} - b_k^{\mathcal{T}}\|_1,$$

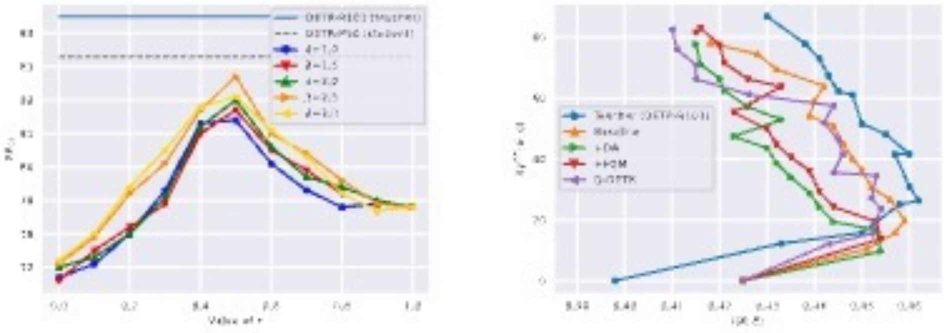
$$\mathcal{L}_{DRD}(\tilde{\mathbf{q}}^{\mathcal{S}^*}, \tilde{\mathbf{q}}^{\mathcal{T}}) = \mathbb{E}[\|\tilde{\mathbf{D}}^{\mathcal{S}^*} - \tilde{\mathbf{D}}^{\mathcal{T}}\|_2],$$

Experiments and Results

Ablation Study

Table 1. Evaluating the components of Q-DETR-R50 on the VOC dataset. #Bits (W-A-Attention) denotes the bit-width of weights, activations, and attention activations. DA denotes the distribution alignment module. FQM denotes foreground-aware query matching.

Method	#Bits	AP_{50}	#Bits	AP_{50}	#Bits	AP_{50}
Real-valued	32-32-32	83.3	-	-	-	-
Baseline	4-4-8	78.0	3-3-8	76.8	2-2-8	69.7
+DA	4-4-8	78.8	3-3-8	78.0	2-2-8	71.6
+FQM	4-4-8	81.5	3-3-8	80.9	2-2-8	74.9
+DA+FQM (Q-DETR)	4-4-8	**82.7**	3-3-8	**82.1**	2-2-8	**76.4**



(a) Effect of τ and λ. (b) Mutual information curves.

Figure 5. (a) We select τ and λ using 4-bit Q-DETR-R50 on VOC. (b) The mutual information curves of $I(X;\mathbf{E})$ and $I(\boldsymbol{y}^{GT};\mathbf{E},\mathbf{q})$ (Eq. 4) on the information plane. The red curves represent the teacher model (DETR-R101). The orange, green, red, and purple lines represent the 4-bit baseline, 4-bit baseline + DA, 4-bit baseline + FQM, and 4-bit baseline + DA + FQM (4-bit Q-DETR).

Experiments and Results

Main Results on VOC

Model	Method	#Bits	AP	AP_{50}	AP_{75}
DETR-R50	Real-valued	32-32-32	59.5	83.3	64.7
	Percentile	8-8-8	54.7	79.2	60.1
	VT-PTQ		57.6	82.3	63.1
	LSQ	4-4-8	49.7	76.9	53.0
	Baseline		51.3	78.0	54.1
	Q-DETR		**57.1**	**82.7**	**61.5**
	LSQ	3-3-8	47.0	75.3	49.1
	Baseline		49.2	76.8	51.8
	Q-DETR		**56.8**	**82.1**	**61.2**
	LSQ	2-2-8	42.6	68.2	44.8
	Baseline		44.0	69.7	45.8
	Q-DETR		**50.7**	**76.4**	**54.1**
SMCA-DETR -R50	Real-valued	32-32-32	56.7	83.7	62.0
	Percentile	8-8-8	54.7	79.2	60.1
	VT-PTQ		55.9	83.0	61.3
	LSQ	4-4-8	49.6	78.6	53.4
	Baseline		50.7	79.5	55.4
	Q-DETR		**56.2**	**83.3**	**61.6**
	LSQ	3-3-8	47.7	76.5	51.7
	Baseline		49.9	77.5	53.6
	Q-DETR		**54.3**	**82.6**	**59.5**
	LSQ	2-2-8	42.3	69.7	44.8
	Baseline		43.9	70.4	46.1
	Q-DETR		**50.2**	**76.7**	**52.6**

For DETR-R50:

- compared with
  the 8-bit PTQ method, our 4-bit Q-DETR achieves a much larger compression ratio than 8-bit VT-PTQ, but with a bit of performance improvement (**82.7% vs. 82.3%**).
- Q-DETR-R50 boosts the performance of 2/3/4-bit baseline by **6.7%, 5.3% and 4.7%** AP with the same architecture and bit-width

For SMCA-DETR-R50:

- Q-DETR with SMCA-DETR-R50 outperforms the 2/3/4-bit Baseline method by **6.3% , 5.1% and 3.8%** on AP50, a large margin.
- Compared with 8-bit post-training quantization methods, our method achieves a significantly higher compression rate and comparable performance

Experiments and Results

Main Results on COCO

Model	Method	#Bits	$Size_{(MB)}$	$OPs_{(G)}$	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
DETR-R50	Real-valued	32-32-32	159.32	85.51	42.0	62.4	44.2	20.5	45.8	61.1
	Percentile	8-8-8	39.83	23.01	38.6	-	-	-	-	-
	VT-PTQ				41.2	-	-	-	-	-
	LSQ				33.3	53.7	33.9	12.8	37.0	51.6
	Baseline	4-4-8	19.92	13.02	34.1	55.3	35.4	14.3	38.0	53.8
	Q-DETR				**39.4**	**60.2**	**41.4**	**17.7**	**43.4**	**59.9**
	LSQ				31.0	52.3	32.1	11.3	33.9	48.5
	Baseline	3-3-8	15.03	7.61	32.3	52.2	32.9	12.3	35.4	50.3
	Q-DETR				**36.1**	**55.9**	**37.5**	**14.6**	**39.4**	**55.2**
	LSQ				24.7	44.6	26.5	6.3	25.3	42.7
	Baseline	2-2-8	10.03	5.32	26.6	46.6	26.5	8.4	28.2	44.4
	Q-DETR				**31.4**	**51.3**	**31.6**	**11.6**	**34.3**	**49.6**
SMCA-DETR-R50	Real-valued	32-32-32	164.75	86.65	41.0	62.2	43.6	21.9	44.3	59.1
	Percentile	8-8-8	41.19	23.66	37.5	58.5	40.1	17.6	39.1	55.9
	VT-PTQ				40.2	61.0	42.6	20.3	42.9	57.7
	LSQ				33.9	55.0	35.0	13.2	37.2	51.4
	Baseline	4-4-8	20.59	13.48	35.0	56.4	36.4	15.6	38.3	52.5
	Q-DETR				**38.3**	**59.7**	**39.8**	**17.7**	**41.7**	**56.8**
	LSQ				30.1	52.6	31.4	11.9	33.4	46.6
	Baseline	3-3-8	15.68	8.05	31.8	53.7	32.6	12.6	35.2	49.8
	Q-DETR				**35.0**	**56.3**	**36.9**	**15.0**	**39.0**	**53.1**
	LSQ				23.9	42.2	24.2	9.4	26.2	37.5
	Baseline	2-2-8	10.84	4.54	25.4	44.3	25.2	8.4	27.2	40.3
	Q-DETR				**30.5**	**51.8**	**31.8**	**12.0**	**33.2**	**48.0**

For DETR-R50:
- Q-DETR-R50 boosts the performance of 2/3/4-bit baseline by **4.8%, 3.8% and 5.1%** AP with the same architecture and bit-width
- 2/3/4-bit Q-DETR-R50 achieves computation acceleration and storage savings by **16.07x/11.23x/6.57x and 15.88x/10.60x/7.99x**, compared to real-valued ones.

For SMCA-DETR-R50:
- 4-bit Q-SMCA-DETR-R50 theoretically accelerates **6.42x** with only a 2.7% performance gap compared with the real-valued counterpart

Conclusion

- This paper introduces a novel method for training quantized DETR (Q-DETR) with knowledge distillation to rectify the query distribution.
- Q-DETR generalizes the information bottleneck (IB) principle and leads a bi-level distribution rectification distillation. We effectively employ a distribution alignment module to solve inner-level and a foreground-aware query matching scheme to solve upper level.
- As a result, Q-DETR significantly boosts performance of low-bit DETR. Extensive experiments show that Q-DETR surpasses state-of-the-arts in DETR quantization.

Thank you for listening