中国图象图形学学会第九期学生会员分享论坛

面向视觉任务的低比特量化方法研究

秦浩桐
ETH Zurich
2024/03/31

基本信息

个人信息

秦浩桐　　1997/07　　神经网络压缩与加速

教育经历

- 2015-2019　北京航空航天大学 BEIHANG UNIVERSITY　计算机学院　本科生
- 2022-2023　ETH zürich　CVL　访问博士生
- 2019-2024　北京航空航天大学 BEIHANG UNIVERSITY　计算机学院 / 沈元学院（实验班）博士生

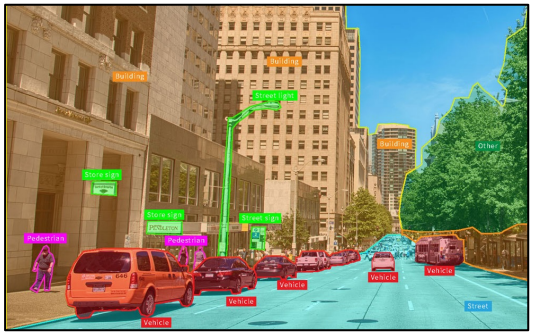
（复杂关键软件环境全国重点实验室　导师：李未院士，刘祥龙教授）

职业经历

- 2018-2019　Microsoft　MSRA　实习研究员（明日之星）
- 2020-2020　Tencent　WXG　实习研究员（犀牛鸟精英人才）
- 2021-2024　ByteDance　AI-Lab　实习研究员
- **2024-至今**　ETH zürich　**PBL**　**博士后研究员**

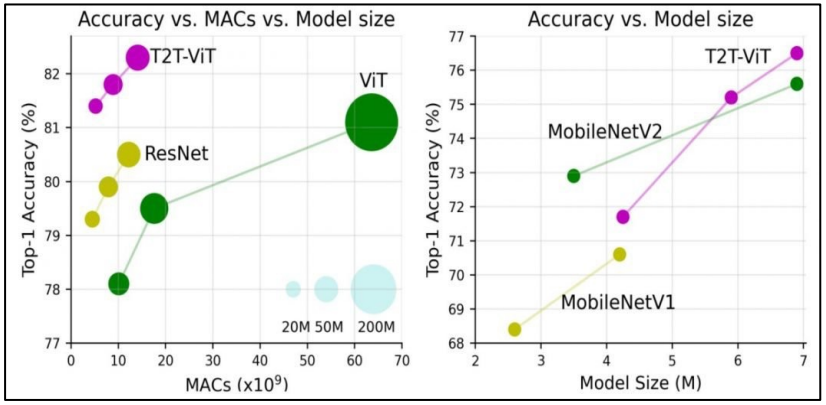
计算机视觉：效率挑战

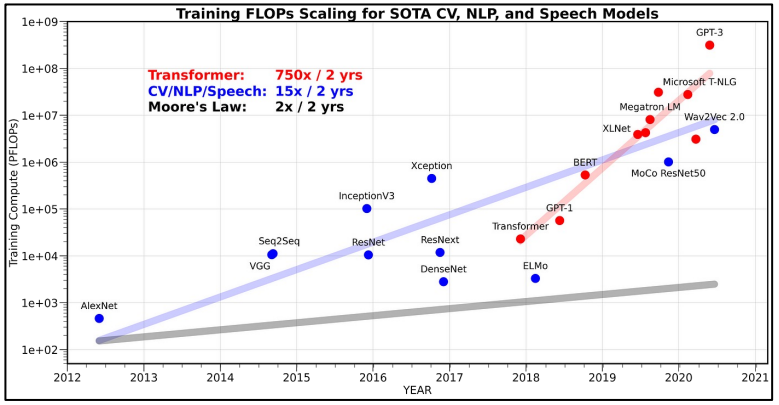
应用潜力



模型尺寸增长与硬件资源限制间的矛盾日益凸显，亟待解决

效率挑战

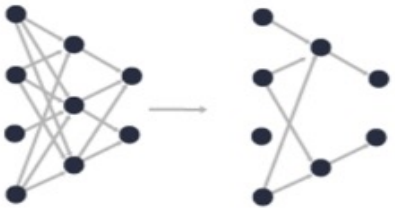




计算机视觉：效率挑战

神经网络压缩与加速方法

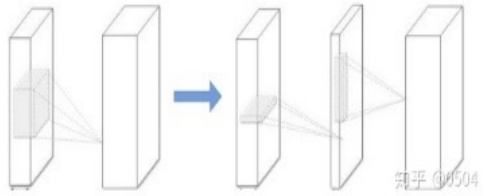
剪枝和参数共享



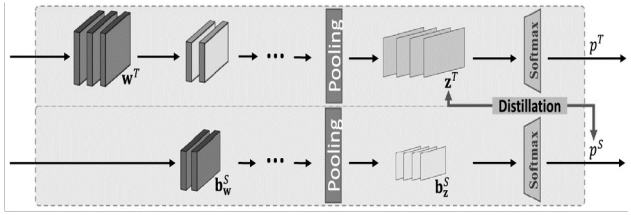
结构化卷积滤波器

$$\mathbf{R} = \mathrm{circ}(\mathbf{r}) := \begin{bmatrix} r_0 & r_{d-1} & \cdots & r_2 & r_1 \\ r_1 & r_0 & r_{d-1} & & r_2 \\ \vdots & r_1 & r_0 & \ddots & \vdots \\ r_{d-2} & & \ddots & \ddots & r_{d-1} \\ r_{d-1} & r_{d-2} & \cdots & r_1 & r_0 \end{bmatrix}.$$

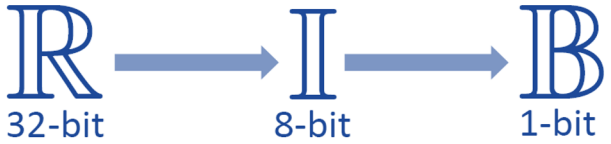
低秩分解



知识蒸馏



低比特量化



Arch	1-bit	4-bit	8-bit
NvGPU-tn	128x	32x	16x
Arm v8.1	5.33x	N/A	1~2x
Arm v8.2	10.67x	N/A	4x
X86 avx2	5.33x	N/A	2~4x
X86 avx512	16x	N/A	4x

常见硬件架构低比特计算的峰值性能对比（相比fp32）

研究内容

模型量化：计算

将模型压缩至有限位宽表达并加速计算

$$\mathbb{R} \xrightarrow[32\text{-bit}]{} \mathbb{I} \xrightarrow[8\text{-bit}]{} \mathbb{B} \in \{-1, +1\} \quad (1\text{-bit})$$

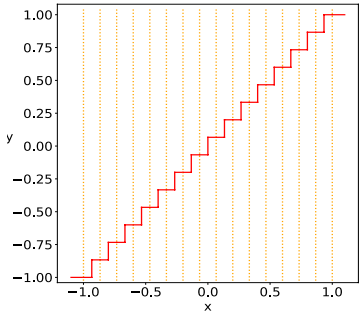
多位量化（精度保持，实用高效）

量化函数：

$$x_{int} = round\left(\frac{x}{\Delta}\right) + z$$

$$x_Q = clamp(0, N_{levels} - 1, x_{int})$$

反量化函数：$x_{float} = (x_Q - z)\Delta$



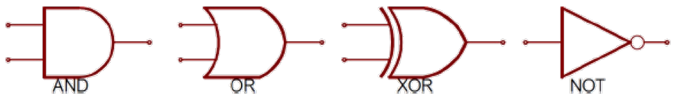
线性量化（整型计算）

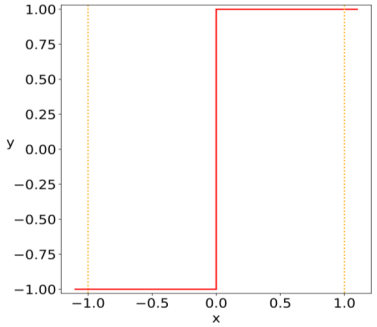
二值量化（超低位宽，节省资源）

量化函数：

$$Q_B(x) = \text{sgn}(x) = \begin{cases} +1, & \text{if } x \geq 0, \\ -1, & \text{otherwise.} \end{cases}$$

按位运算：





二值量化（按位运算）

研究内容

模型量化：管线

原始数据

训练硬件

偏置
(32-bit)

求和

训练
数据集

权重
(32-bit 代理)

激活
(32-bit 代理)

损失函数

伪量化模型
(32-bit)

量化

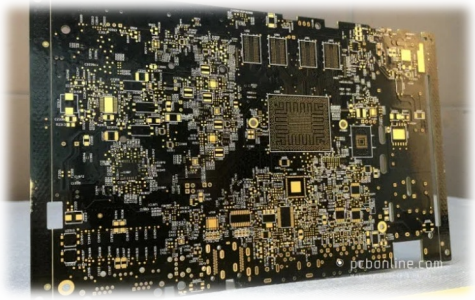
测试
数据集

伪量化

最终推理结果

量化模型
(1~8-bit)

模型部署

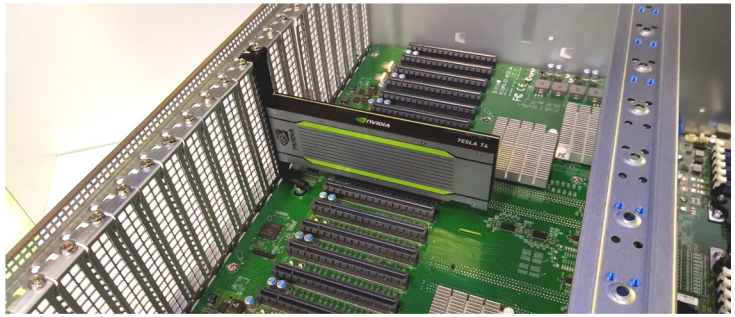


推理硬件

研究内容

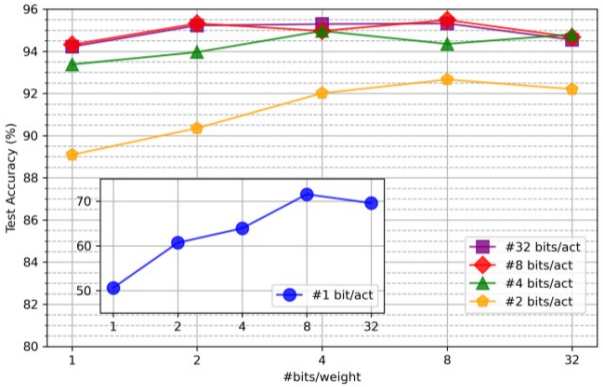
挑战1: 面向 有限训练资源 的视觉模型量化

隐私场景



有限资源

挑战2: 面向 有限推理资源 的视觉模型量化



95% (4-bit)

⇩

50% (1-bit)

精度陡降

超低算力

关键挑战一：有限训练资源量化

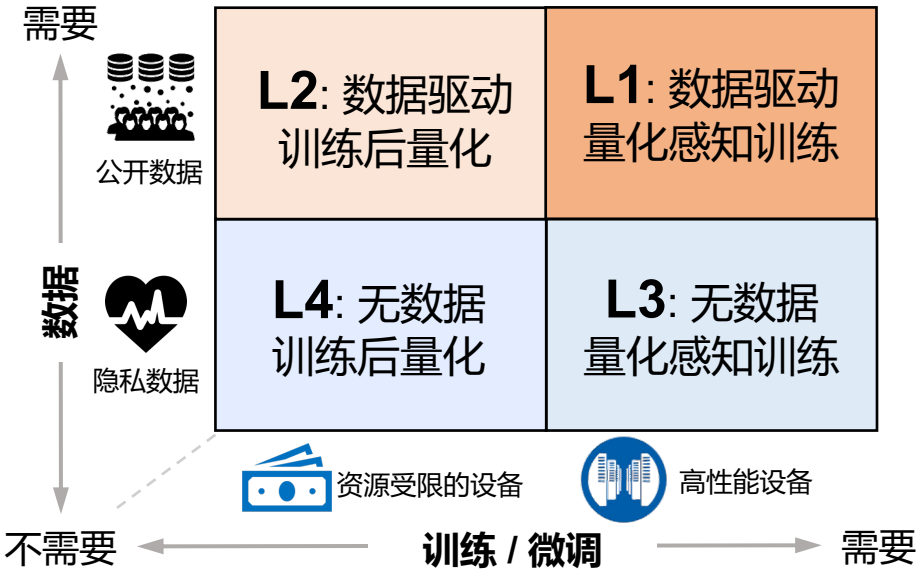
多位量化（2-8bit）：落地实用的低比特量化，面向更广场景

- 部分位宽已高度实用化，几乎不损失精度
- 对各种深度学习模型、应用适用广泛

$$\mathbb{R} \longrightarrow \mathbb{I}$$

32-bit　2/4/8-bit

- **挑战**
  - **无数据、低算力**等限制场景下的视觉模型量化生产和优化

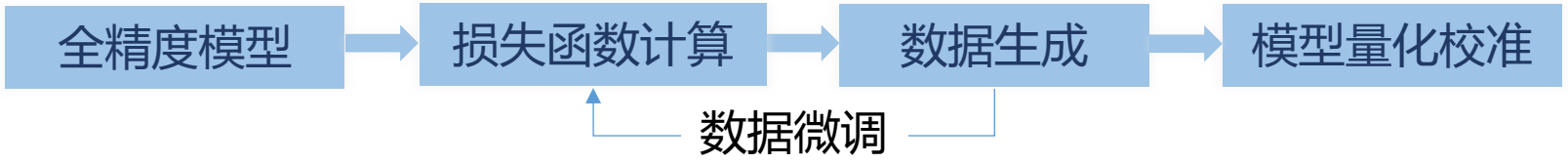


Post Training Quantization (PTQ)　　Quantization Aware Training (QAT)

Level-4 量化：面向无数据量化的样本生成

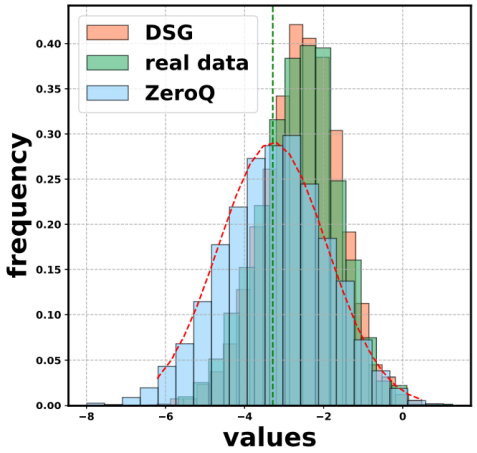
问题：如何弥合真实与合成数据间的量化差距

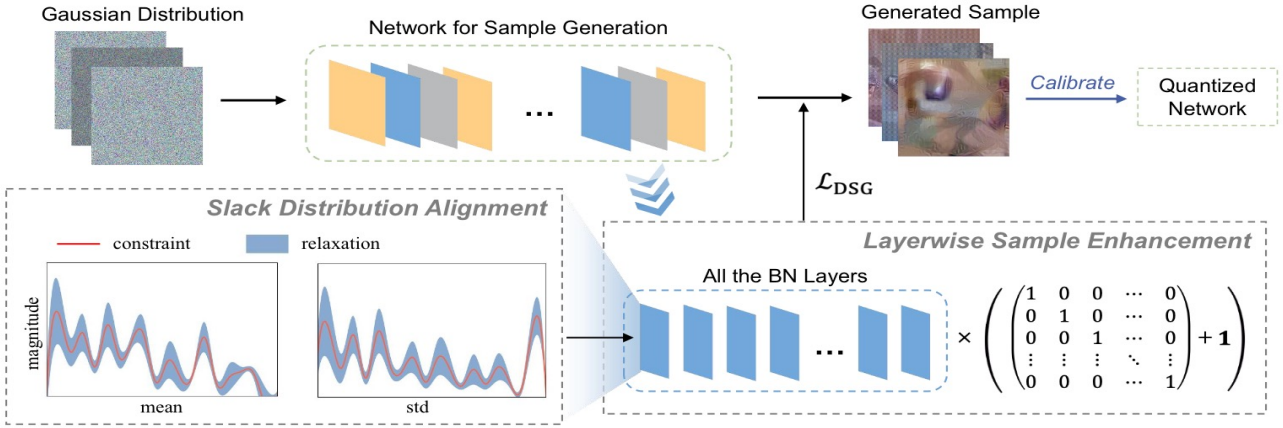
思路：设计多样化数据的生成式无数据量化



无数据场景性能低 → 量化依赖真实样本

数据多样性是合成样本与真实样本间的重要区别





其中SDA和LSE在统计层面增强数据的分布和样本多样性

CVPR 2021 (CCF A会议, Oral)

Level-4 量化：面向无数据量化的样本生成

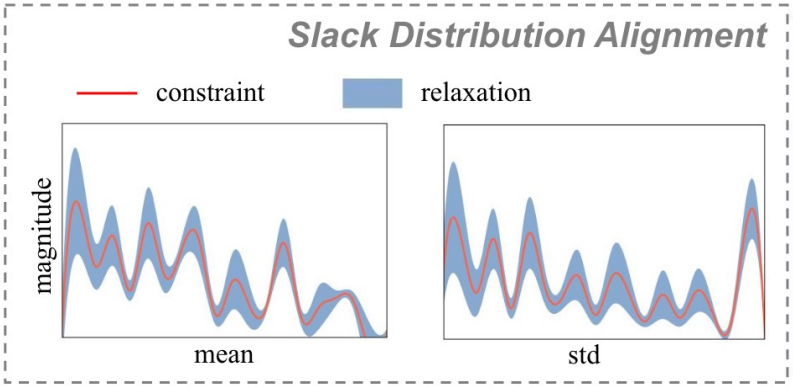
问题： 如何弥合真实与合成数据间的量化差距

思路： 设计多样化数据的生成式无数据量化

松弛分布对齐

$$l_{\mathrm{SDA}_i} = \left\|\max\left(\left|\tilde{\boldsymbol{\mu}}_i^s - \boldsymbol{\mu}_i\right| - \delta_i, 0\right)\right\|_2^2 + \left\|\max\left(\left|\tilde{\boldsymbol{\sigma}}_i^s - \boldsymbol{\sigma}_i\right| - \gamma_i, 0\right)\right\|_2^2$$

$$\delta_i = \left|\tilde{\boldsymbol{\mu}}_i^0 - \boldsymbol{\mu}_i\right|_\epsilon \quad \gamma_i = \left|\tilde{\boldsymbol{\sigma}}_i^0 - \boldsymbol{\sigma}_i\right|_\epsilon$$



逐层样本增强

Sample *1* → $\min_{x^r} \sum_{i=0}^{L} \|\tilde{\mu}_i^r - \mu_i\|_2^2 + \|\tilde{\sigma}_i^r - \sigma_i\|_2^2, + (\|\tilde{\mu}_0^r - \mu_0\|_2^2 + \|\tilde{\sigma}_0^r - \sigma_0\|_2^2)$

Sample *2* → $\min_{x^r} \sum_{i=0}^{L} \|\tilde{\mu}_i^r - \mu_i\|_2^2 + \|\tilde{\sigma}_i^r - \sigma_i\|_2^2, + (\|\tilde{\mu}_1^r - \mu_1\|_2^2 + \|\tilde{\sigma}_1^r - \sigma_1\|_2^2)$

...

Sample *n* → $\min_{x^r} \sum_{i=0}^{L} \|\tilde{\mu}_i^r - \mu_i\|_2^2 + \|\tilde{\sigma}_i^r - \sigma_i\|_2^2, + (\|\tilde{\mu}_n^r - \mu_n\|_2^2 + \|\tilde{\sigma}_n^r - \sigma_n\|_2^2)$

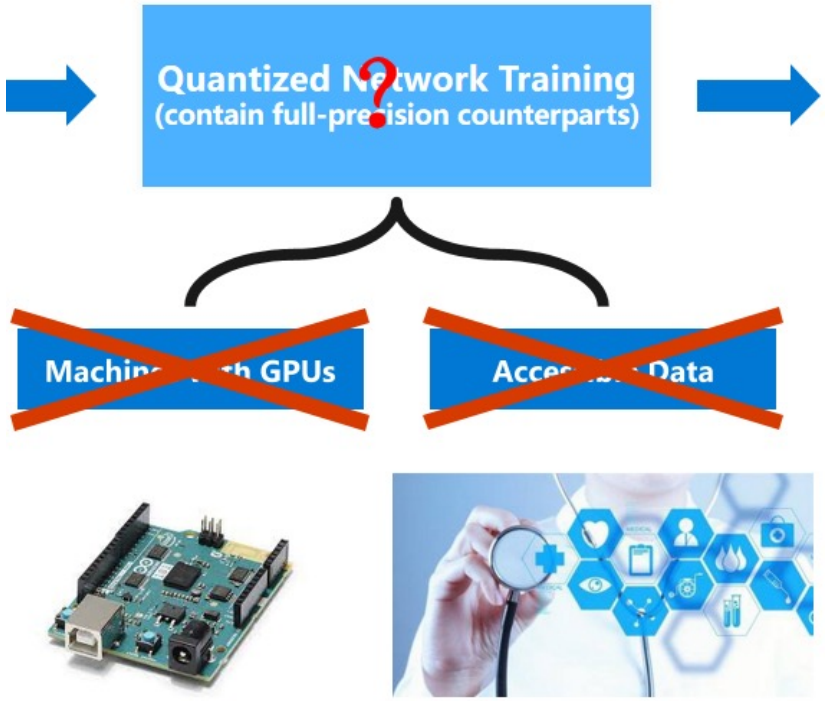
CVPR 2021 (CCF A会议, Oral)

Level-4 量化：面向无数据量化的样本生成

卷积模型无数据量化收敛精度下 位宽最低

方法	机构	精度	文献
DSG	**北京航空航天大学，商汤**	**34.53%**	**CVPR 2021**
ZeroQ	北京大学	26.04%	CVPR 2020
KL	首尔国立大学	16.27%	ICML 2015
MSE	华盛顿大学	15.08%	ECCV 2018
ACIQ	英特尔	7.19%	ICLR 2018
DFQ	高通	0.10%	ICCV 2019

INT4量化精度比较 (ResNet18, ImageNet)



相对于**英特尔**、**高通**等成果，多个任务、结构上INT4**精度收敛**

面向**边缘设备**、**隐私保护**等低资源场景下量化模型生产

CVPR 2021 (CCF A会议, Oral)

Level-3 量化：面向无数据量化的量化训练

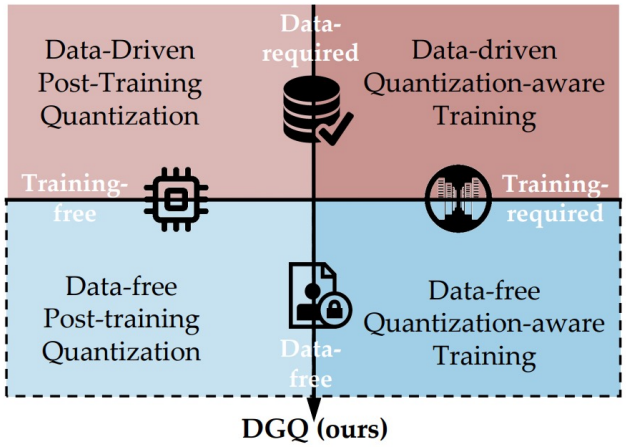
问题：如何改善量化训练中无数据量化精度

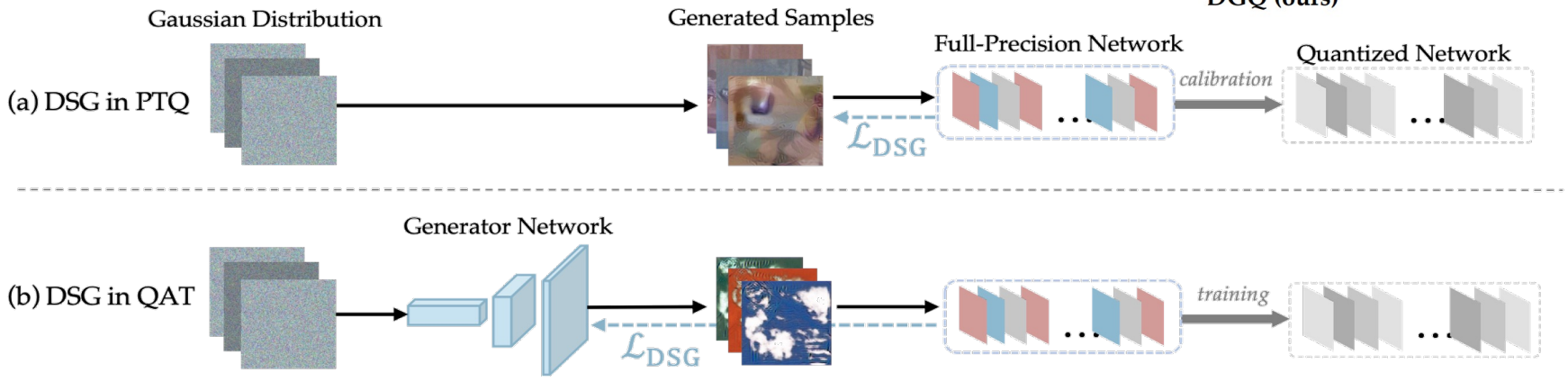
思路：对合成数据和量化模型进行联合优化

生成式无数据量化：
从离线校准迈向无数据量化训练

$$\mathcal{L}_{\text{DSG-PTQ}} = \left(\mathbf{X}_{\text{LSE}} \boldsymbol{l}_{\text{SDA}}\right)^T J_{N,1} + l_{\text{SCI}}.$$

$$\mathcal{L}_{\text{DSG-QAT}} = \left(\mathbf{X}_{\text{LSE}} \boldsymbol{l}_{\text{SDA}}\right)^T J_{N,1} + \boldsymbol{l}_{\text{SCI}}^T J_{N,1} + \mathcal{L}_{\text{CE}}^{\text{G}}(\text{G})$$



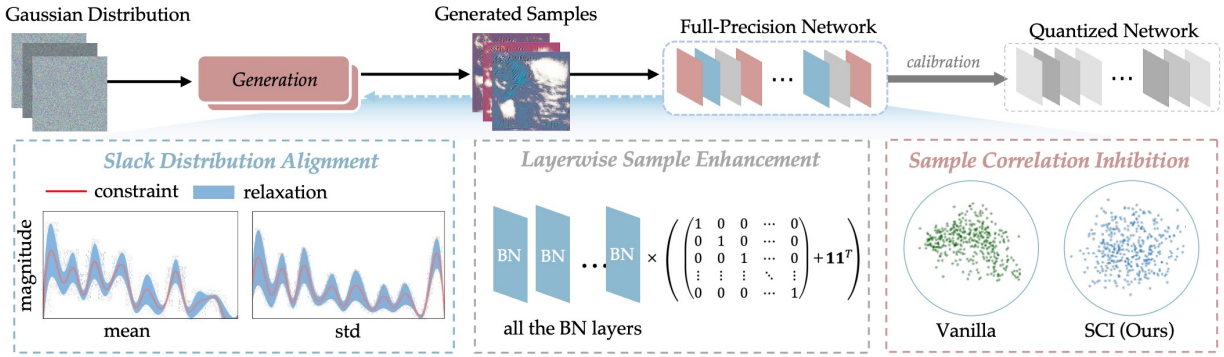


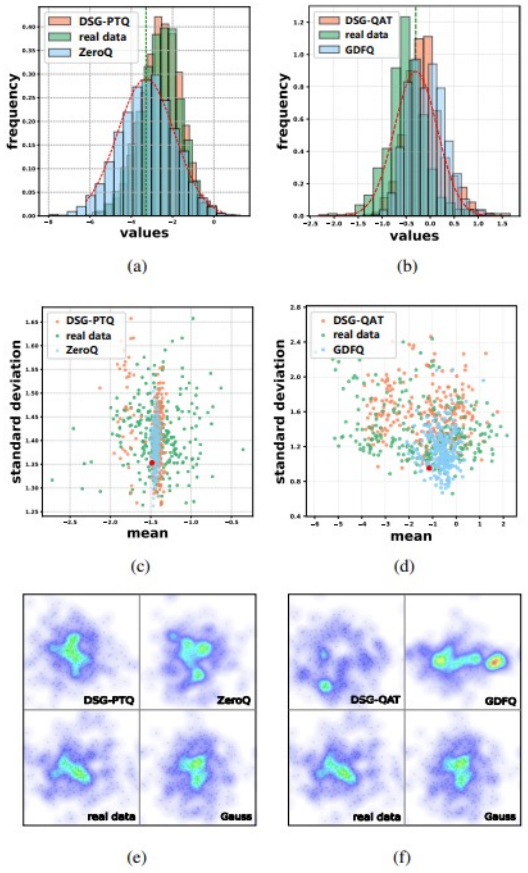
IEEE TPAMI 2023 (CCF A期刊)

Level-3 量化：面向无数据量化的量化训练

问题：如何改善量化训练中无数据量化精度

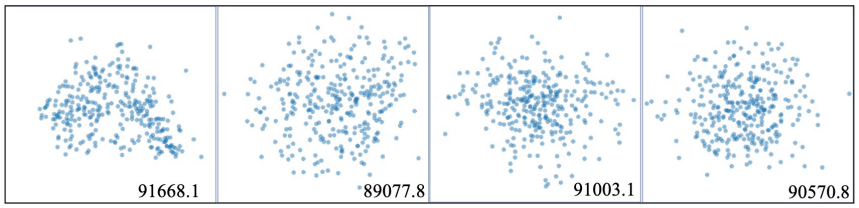
思路：对合成数据和量化模型进行联合优化





优化生成数据 + 量化模型：（统计层面外）通过SCI在空间层面增强数据的多样性

$$l_{\mathrm{SCI}} = \max\left(\sum_i \left(\left\|\lambda_{\mathbf{f}}^i - \lambda_{\mathbf{r}}^i\right\|_2 - \hat{\lambda}_{\mathbf{f}}^i \cos\left(v_{\mathbf{f}}^i, v_{\mathbf{r}}^i\right)\right), 0\right)$$

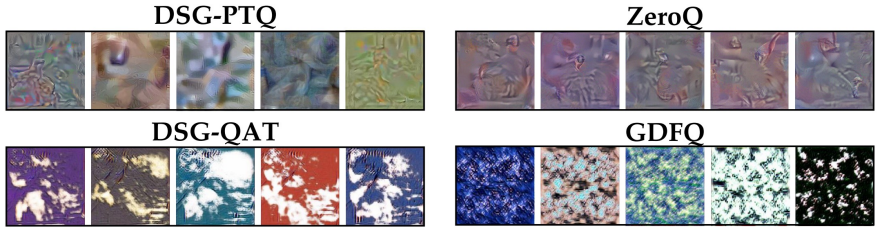


(a) GDFQ (b) ours (c) real data (d) Gaussian

IEEE TPAMI 2023 (CCF A期刊)

Level-3 量化：面向无数据量化的量化训练

无数据量化样本生成多样性 理论探索



Theorem 1. *Given a set of all possible input spaces* $\mathbf{X} = \{\mathcal{X}_1, \mathcal{X}_2, \cdots\}$, *where each element* $\mathcal{X}_i$ *has a corresponding overall density* ρ_i, *volume* V_i, *and mass* $M_i = \rho_i \cdot V_i$. *Additionally,* $\mathcal{X}_i$ *is comprised of several regions* $\{\mathcal{R}_1^i, \cdots, \mathcal{R}_{K^i}^i\}$ *with unknown, limited* $K^i \geq 2$. *The region* $\mathcal{R}_j^i, 1 \leq j \leq K^i$ *also have its own density* ρ_j^i, *volume* V_j^i, *and mass* M_j^i. *Let us consider a sample set* $\mathbf{x}^s = \{x_1^s, \cdots, x_N^s\} \subset \mathcal{X}^*$, *where* $\mathcal{X}^* = \mathbb{E}(\mathbf{X})$ *represents the latent input space with mass* M^*. *And* $\mathcal{X}^* = \mathcal{R}_1^* \cup \cdots \cup \mathcal{R}_{K^*}^*$, *where the mass of* $\mathcal{R}_j^*$ *is* M_j^* *and* $K^* = \max(K^1, K^2, \cdots)$. *If* $\mathbf{x}^s$ *satisfies the condition that* $\forall x_i^s \in \mathbf{x}^s$, $p(x_i^s \in \mathcal{R}_j^*) = \frac{M_j^*}{M^*}$, *the information reflected from* $\mathbf{X}$ *by the sample set* $\mathbf{x}^s$ *will be maximized in mathematical expectation, where* $M^* = \sum_{j=1}^{K^*} M_j^*$.

	Baseline	–	–	32	32	77.72
ResNet50	Real Data	✗	✗	4	4	70.27
	GDFQ	✓	✗	4	4	55.65
	RVQuant	✗	✗	4	4	64.90
	ZAQ	✓	✗	4	4	70.06
	DSG (Ours)	✓	✗	4	4	**71.96**
	Real Data	✗	✗	6	6	77.56
	GDFQ	✓	✗	6	6	76.59
	ZS-CGAN	✓	✗	6	6	76.82
	DSG (Ours)	✓	✗	6	6	**77.25**
	Real Data	✗	✗	8	8	77.66
	GDFQ	✓	✗	8	8	77.51
	DSG (Ours)	✓	✗	8	8	**77.64**

首次给出了样本丰富性在生成式量化中重要性的**理论证明**

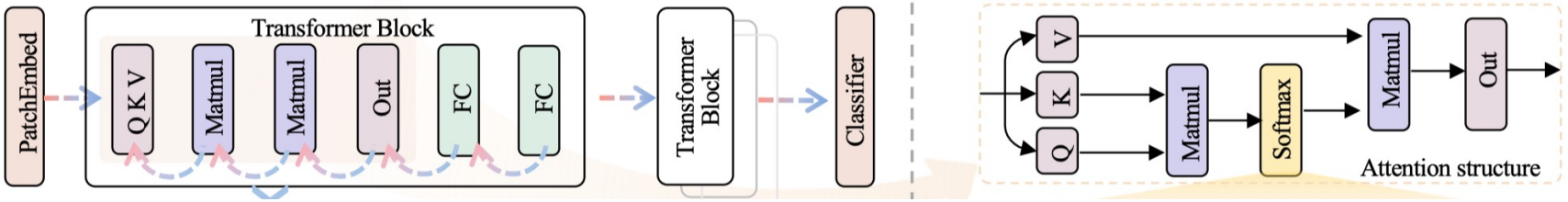
当前无数据量化**SOTA**性能
INT4 ResNet50仅损失6%

IEEE TPAMI 2023 (CCF A期刊)

Level-2 量化：面向训练后量化的校准修正

问题： 离线量化中注意力结构中的特征分布独特

思路： 修正校准度量和校准方式，减小量化误差

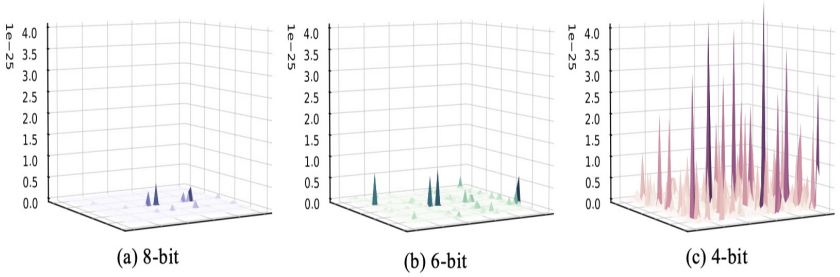


结构特点： （1）按块堆叠 （2）注意力机制

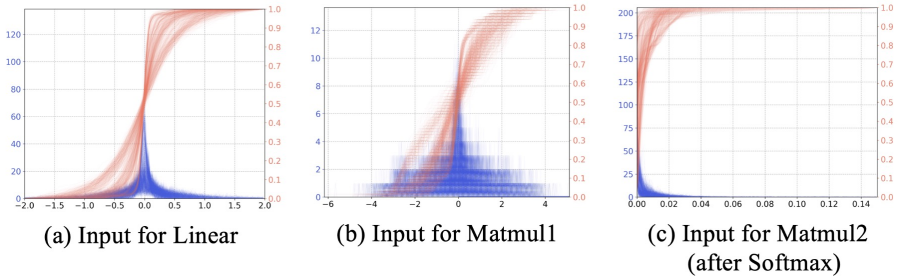
离线量化：

$$\mathbb{E}[\mathcal{L}(\hat{\mathbf{x}})] - \mathbb{E}[\mathcal{L}(\mathbf{x})] \approx \epsilon^T \bar{g}^{(\mathbf{x})} + \frac{1}{2}\epsilon^T \bar{H}^{(\mathbf{x})}\epsilon, \quad \epsilon^T \bar{H}^{(\mathbf{x})}\epsilon \approx (\hat{O}-O)^T \bar{H}^{(O)}(\hat{O}-O).$$

超低比特量化损失显著



softmax输出分布不对称不平衡



ACM MM 2022 (CCF A会议)

Level-2 量化：面向训练后量化的校准修正

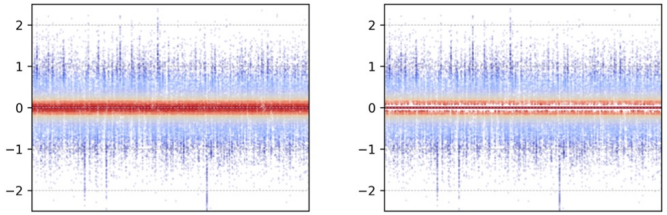
问题： 离线量化中保持注意力结构特性并准确优化

思路： 修正校准度量和校准方式，减小量化误差

超低比特量化损失显著

逐块底部消除策略

从全局视角减小关键误差



(a) Original (b) Bottom-elimination

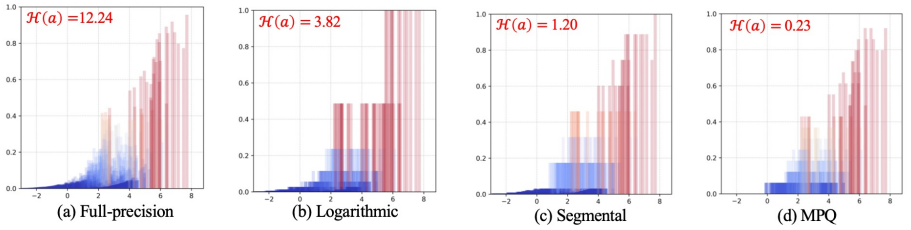
$$\sigma_\gamma^b = \sigma^b[\sigma^b]_\gamma, \qquad [\sigma^b]_\gamma = \begin{cases} 1, \text{where } \sigma^b < |\sigma^b|_\gamma, \\ 0, \text{otherwise}, \end{cases}$$

$$\min_{\Delta} \mathbb{E}\left[\left(\sigma^b - \sigma_\gamma^b\right)^T \operatorname{diag}\left(\left(\frac{\partial \mathcal{L}}{\partial O_1^b}\right)^2, \ldots, \left(\frac{\partial \mathcal{L}}{\partial O_{|O^b|}^b}\right)^2\right)\left(\sigma^b - \sigma_\gamma^b\right)\right]$$

softmax输出分布不对称不平衡

马太效应保持的softmax量化

保持量化后注意力结构的幂率分布



(a) Full-precision (b) Logarithmic (c) Segmental (d) MPQ

$$\text{softmax}(\mathbf{x})_i = \frac{e^{\beta x_i}}{\sum_{j=1}^{K} e^{\beta x_j}}, \text{for } i = 1, \ldots, K,$$

$$\hat{\mathbf{x}_s} = \text{clamp}\left(\left\lfloor \frac{\text{softmax}(\mathbf{x})}{\Delta} \right\rfloor, 0, 2^k - 1\right), \quad \Delta = \frac{\max(\text{softmax}(\mathbf{x}))}{2^k - 1}$$

ACM MM 2022 (CCF A会议)

Level-2 量化：面向训练后量化的校准修正

离线量化中对各类ViT模型 适用最广

方法	机构	精度	文献
APQ-ViT	**北京航空航天大学**	**75.64%**	**ACM MM 2021**
PTQ4ViT	北京大学	75.58%	NeurIPS 2021
FQ-ViT	旷视	58.87%	IJCAI 2022

量化模型精度比较 (ViT-S, ImageNet)

方法	机构	精度	文献
APQ-ViT	**北京航空航天大学**	**0.458**	**ACM MM 2021**
PTQ4ViT	北京大学	0.271	NeurIPS 2021

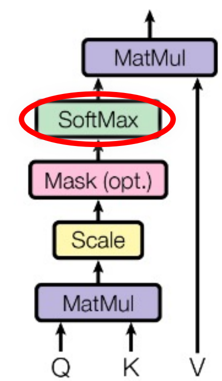
量化模型精度比较 (MaskRCNN-SwinT, COCO)

Method	#bit(W/A)	ViT-T	ViT-S	ViT-S/32	ViT-B	DeiT-T	DeiT-S	DeiT-B	Swin-S	Swin-B	Swin-B/384
Full-precision	32/32	75.47	81.39	75.99	84.54	72.21	79.85	81.80	83.23	85.27	86.44
FQ-ViT	4/4	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
PTQ4ViT	4/4	17.45	42.57	35.09	30.69	36.96	34.08	64.39	76.09	74.02	78.84
APQ-ViT (Ours)	4/4	**17.56**	**47.95**	**41.53**	**41.41**	**47.94**	**43.55**	**67.48**	**77.15**	**76.48**	**80.84**

Method	#bit(W/A)	Mask RCNN Swin-T		Mask RCNN Swin-S		Cascade Mask RCNN Swin-T		Cascade Mask RCNN Swin-S		Cascade Mask RCNN Swin-B	
		AP^{box}	AP^{mask}	AP^{box}	AP^{mask}	AP^{box}	AP^{mask}	AP^{box}	AP^{mask}	AP^{box}	AP^{mask}
Full-precision	32/32	46.0	41.6	48.5	43.3	50.4	43.7	51.9	45.0	51.9	45.0
BasePTQ	4/4	0.9	0.9	12.6	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	4.0	3.7
PTQ4ViT	4/4	6.9	7.0	26.7	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	10.6	9.3
APQ-ViT (Ours)	4/4	23.7	22.6	44.7	40.1	27.2	24.4	47.7	41.1	47.6	41.5

15个ViT及变体
2个大型视觉任务

#bit(W/A)	Quantizer	ViT-S	DeiT-S	DeiT-B
Full-precision	-	81.39	79.85	81.80
4/4	Log	44.72	21.59	60.91
	Segmental	37.70	22.31	60.02
	MPQ	**47.95**	**43.55**	**67.48**
6/6	Log	78.94	77.57	80.34
	Segmental	78.67	76.64	80.37
	MPQ	**79.10**	**77.76**	**80.42**
8/8	Log	81.13	79.76	81.70
	Segmental	81.00	79.47	81.70
	MPQ	**81.25**	**79.78**	**81.72**



适用改进Transformer、卷积网络等**多种结构**离线量化

ACM MM 2022 (CCF A会议)

Level-1 量化：面向量化训练的表征恢复

问题： 在量化训练中恢复视觉模型的细粒度高频表征

思路： 改善模型中量化算子和模型架构的学习能力

超分任务特点： 量化模型中的表征细节质量要求高

超低比特性能损失明显

模型参数量化后高频信息难以恢复

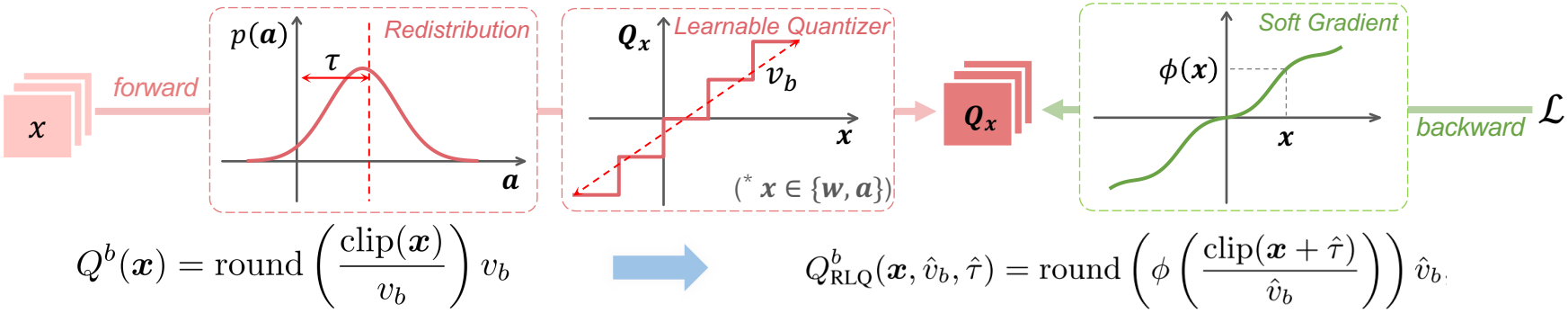
NeurIPS 2023 (CCF A会议, Spotlight)

Level-1 量化：面向量化训练的表征恢复

问题： 在量化训练中恢复视觉模型的细粒度高频表征

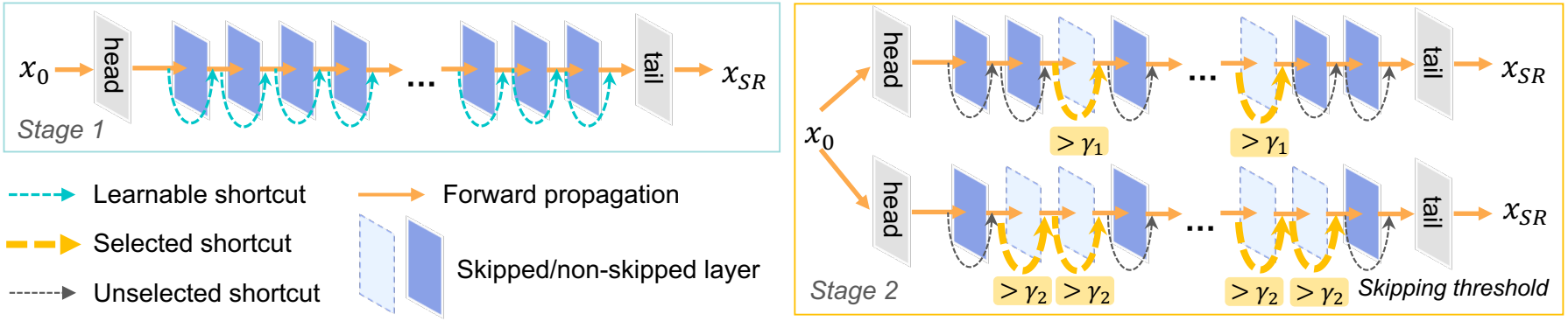
思路： 改善模型中量化算子和模型架构的学习能力

重分配驱动可学习量化器



$$Q^b(\boldsymbol{x}) = \text{round}\left(\frac{\text{clip}(\boldsymbol{x})}{v_b}\right) v_b \quad \Rightarrow \quad Q^b_{\text{RLQ}}(\boldsymbol{x}, \hat{v}_b, \hat{\tau}) = \text{round}\left(\phi\left(\frac{\text{clip}(\boldsymbol{x}+\hat{\tau})}{\hat{v}_b}\right)\right)\hat{v}_b$$

深度动态量化架构



NeurIPS 2023 (CCF A会议, Spotlight)

Level-1 量化：面向量化训练的表征恢复

超分任务CNN和Transformer架构上 最优性能

Method	Scale	#Bit (w/a)	Set5		Set14		B100		Urban100		Manga109	
			PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	×2	-/-	33.66	0.9299	30.24	0.8688	29.56	0.8431	26.88	0.8403	30.80	0.9339
SRResNet [21]	×2	32/32	38.00	0.9605	33.59	0.9171	32.19	0.8997	32.11	0.9282	38.56	0.9770
SwinIR_S [26]	×2	32/32	38.14	0.9611	33.86	0.9206	32.31	0.9012	32.76	0.9340	39.12	0.9783
DoReFa [44]	×2	8/8	37.32	0.9520	32.90	0.8680	31.69	0.8504	30.32	0.8800	37.01	0.9450
CADyQ [11]	×2	8/8	37.79	0.9590	33.37	0.9150	32.02	0.8980	31.53	0.9230	38.06	0.9760
DoReFa [44]	×2	4/4	37.31	0.9510	32.48	0.9091	31.64	0.8901	30.18	0.8780	36.95	0.9440
PAMS [23]	×2	4/4	37.67	0.9588	33.19	0.9146	31.90	0.8966	31.10	0.9194	37.62	0.9400
CADyQ [11]	×2	4/4	37.58	0.9580	33.14	0.9140	31.87	0.8960	30.94	0.9170	37.31	0.9740
QuantSR-C (ours)	×2	4/4	37.80	0.9597	33.35	0.9158	32.04	0.8979	31.46	0.9221	38.25	0.9762
QuantSR-T (ours)	×2	4/4	38.10	0.9604	33.65	0.9186	32.21	0.8998	32.20	0.9295	38.85	0.9774
DoReFa [44]	×2	2/2	36.91	0.9470	32.55	0.9071	31.41	0.8868	29.60	0.8740	36.132	0.9410
PAMS [23]	×2	2/2	34.04	0.8270	30.91	0.8751	30.11	0.8592	27.57	0.8400	31.79	0.9110
CADyQ [11]	×2	2/2	19.44	0.5610	18.51	0.4810	19.70	0.4760	17.97	0.4550	17.346	0.5830
QuantSR-C (ours)	×2	2/2	37.57	0.9589	33.09	0.9136	31.84	0.8954	30.77	0.9149	37.60	0.9745
QuantSR-T (ours)	×2	2/2	37.55	0.9587	33.12	0.9143	31.89	0.8958	30.96	0.9172	37.61	0.9745
Bicubic	×4	-/-	28.42	0.8104	26.00	0.7027	25.96	0.6675	23.14	0.6577	24.89	0.7866
SRResNet [21]	×4	32/32	32.16	0.8951	28.60	0.7822	27.58	0.7364	26.11	0.7870	30.46	0.9089
SwinIR_S [26]	×4	32/32	32.44	0.8976	28.77	0.7858	27.69	0.7406	26.47	0.7980	30.92	0.9151
DoReFa [44]	×4	4/4	29.57	0.8369	26.82	0.7352	26.47	0.6971	23.75	0.6898	27.89	0.8634
PAMS [23]	×4	4/4	31.59	0.8851	28.20	0.7725	27.32	0.7220	25.32	0.7624	28.86	0.8805
CADyQ [11]	×4	4/4	31.48	0.8830	28.05	0.7690	27.21	0.7240	25.09	0.7520	28.82	0.8840
QuantSR-C (ours)	×4	4/4	32.00	0.8924	28.50	0.7799	27.52	0.7342	25.88	0.7807	30.15	0.9040
QuantSR-T (ours)	×4	4/4	32.18	0.8941	28.63	0.7822	27.59	0.7367	26.11	0.7871	30.49	0.9087
DoReFa [44]	×4	2/2	30.54	0.8610	27.50	0.7538	26.90	0.7098	24.44	0.7242	27.31	0.8502
PAMS [23]	×4	2/2	29.20	0.8239	26.61	0.7273	26.36	0.6934	23.58	0.6812	25.59	0.8012
CADyQ [11]	×4	2/2	19.67	0.5380	19.30	0.4740	19.80	0.4620	17.97	0.4360	17.30	0.5640
QuantSR-C (ours)	×4	2/2	31.30	0.8819	28.08	0.7694	27.23	0.7246	25.13	0.7537	28.81	0.8844
QuantSR-T (ours)	×4	2/2	31.53	0.8845	28.16	0.7715	27.28	0.7274	25.26	0.7609	29.06	0.8898

精度提升

超低位宽

困难尺度

架构通用

NeurIPS 2023 (CCF A会议, Spotlight)

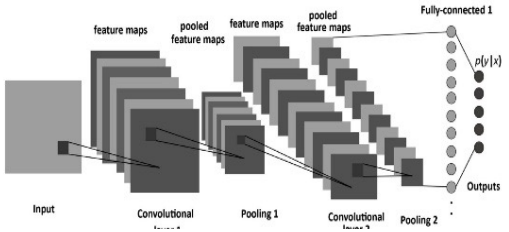
关键挑战二：有限推理资源量化

二值量化（1-bit）：极低位宽的量化训练，面向更高精度

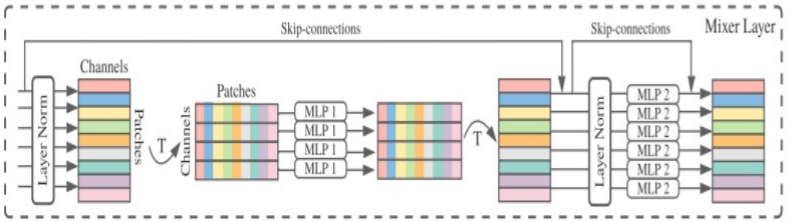
- 量化技术的终极形态，相同结构下的最小存储空间
- 适用于按位运算，充分利用硬件加速潜力

- **挑战**
  - **精度损失巨大**是阻碍二值化落地的最关键原因

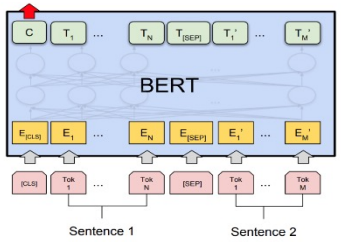
实验观察：结构是影响极低位宽量化精度的主要因素之一



卷积结构 (ResNet, VGG, ...)



全连接结构 (PointNet, ...)



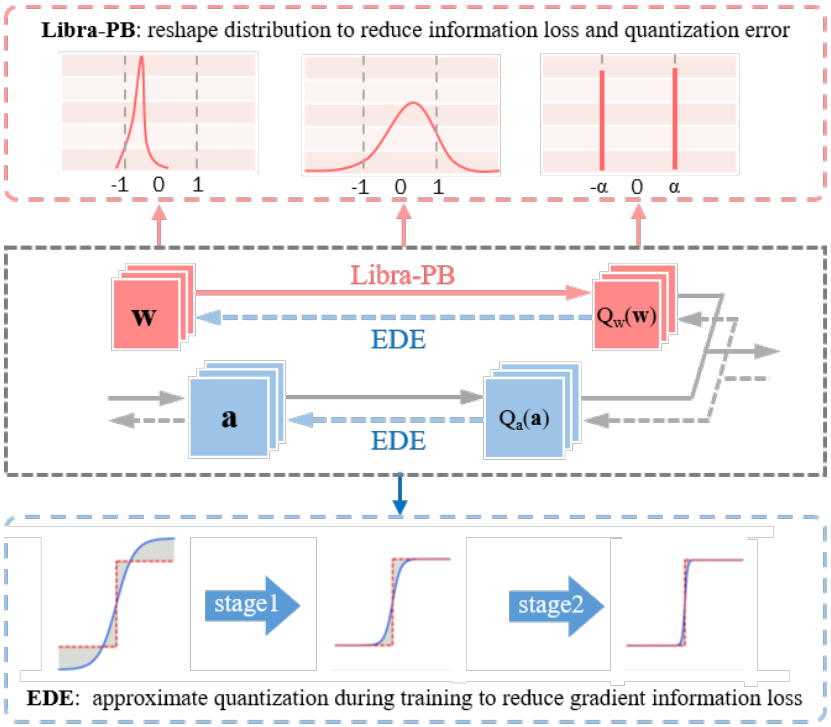
Transformer 结构 (DiT, ViT,...)

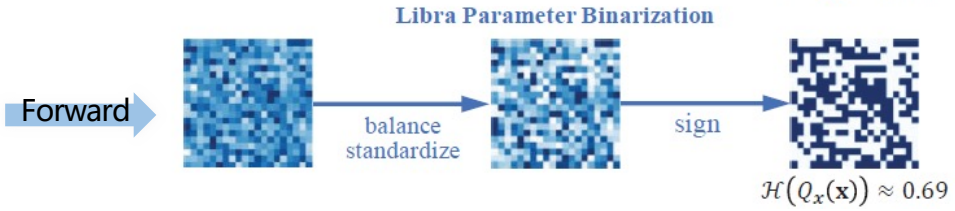
典型架构如何实现高准确率的极低位宽量化训练？

通用算子量化：前后向量化感知信息保持 (图像分类、检测)

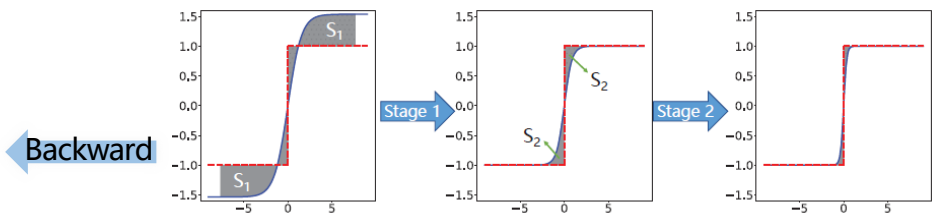
问题： 在对卷积表征二值化的同时降低其信息损失

思路： 提出二值算子的前后向信息保留机制





前向传播: 对权重进行平衡二值量化



反向传播: 渐进两阶段梯度近似方法

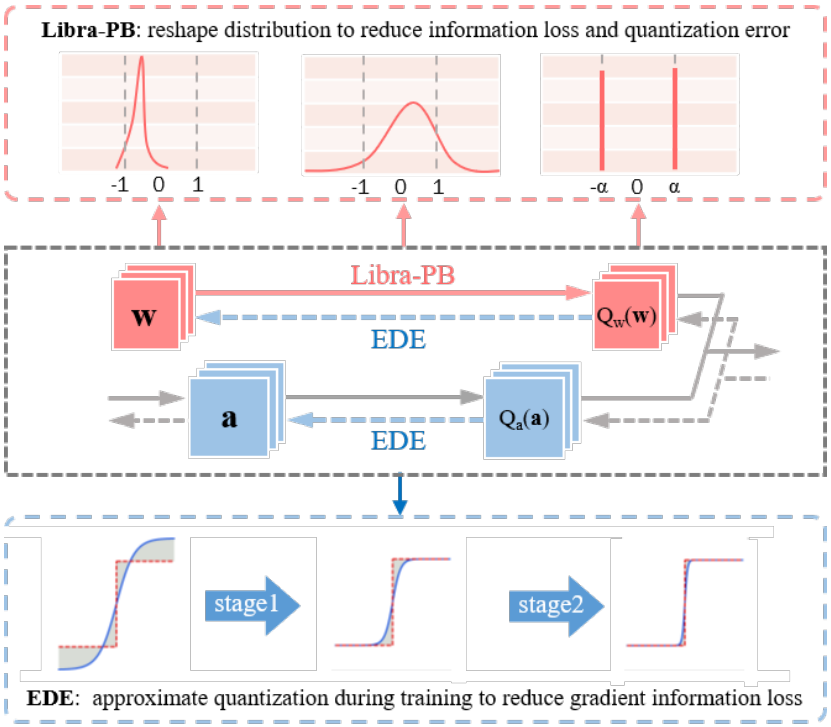
IR-Net： 首次从信息的角度研究二值神经网络的训练过程并提高网络性能

CVPR 2020 (CCF A会议) IJCV 2023 (CCF A期刊)

通用算子量化：前后向量化感知信息保持 (图像分类、检测)

问题： 在对卷积表征二值化的同时降低其信息损失

思路： 提出二值算子的前后向信息保留机制



信息视角：二值参数的熵最大化

$$f(b) = \begin{cases} p, & if\ b = +1 \\ 1-p, & if\ b = -1, \end{cases}$$

$$\mathcal{H}(Q_x(\mathbf{x})) = \mathcal{H}(\mathbf{B_x}) = -p\ln(p) - (1-p)\ln(1-p).$$

$$\min J(Q_x(\mathbf{x})) - \lambda\mathcal{H}(Q_x(\mathbf{x})).$$

$$\hat{\mathbf{w}}_{\text{std}} = \frac{\hat{\mathbf{w}}}{\sigma(\hat{\mathbf{w}})}, \quad \hat{\mathbf{w}} = \mathbf{w} - \overline{\mathbf{w}}.$$

$$\mathbb{E}[z] = Q_w(\hat{\mathbf{w}}_{\text{std}})^\top \mathbb{E}[Q_a(\mathbf{a})] = Q_w(\hat{\mathbf{w}}_{\text{std}})^\top \mu\mathbf{1}.$$

IR-Net：首次从信息的角度研究二值神经网络的训练过程并提高网络性能

CVPR 2020 (CCF A会议) IJCV 2023 (CCF A期刊)

通用算子量化：前后向量化感知信息保持 (图像分类、检测)

相对同期学术界和工业界相关工作 表征最准

方法	机构	精度	文献
DIR-Net	**北京航空航天大学**	**66.50%**	**IJCV 2022**
IR-Net	北京航空航天大学	58.10%	CVPR 2020
ReActNet	卡耐基梅隆大学	65.90%	ECCV 2020
Si-BNN	中国科学院自动化研究所	59.70%	AAAI 2020
ABC-Net	大疆科技	42.70%	NeurIPS 2017
XNOR-Net	华盛顿大学	51.20%	ECCV 2016
Dorefa	旷视科技	53.40%	TensorPack 2016

二值模型精度比较 (ResNet-18, ImageNet)

国际上首次报告硬件部署性能

商汤2018青年基金（亚洲**10**项）

Method	Bit-width (W/A)	Size (Mb)	Time (ms)
FP	32/32	46.77	1418.94
NCNN	8/8	–	935.51
DSQ	2/2	–	551.22
Ours (without bit-shift scales)	1/1	**4.20**	**252.16**
Ours	1/1	**4.21**	**261.98**

	DSQ 2-bit	NCNN 8-bit [31]
time (ms)	**551.22**	949.63

相对于**CMU**、**大疆**等成果取得多个任务**新SOTA**

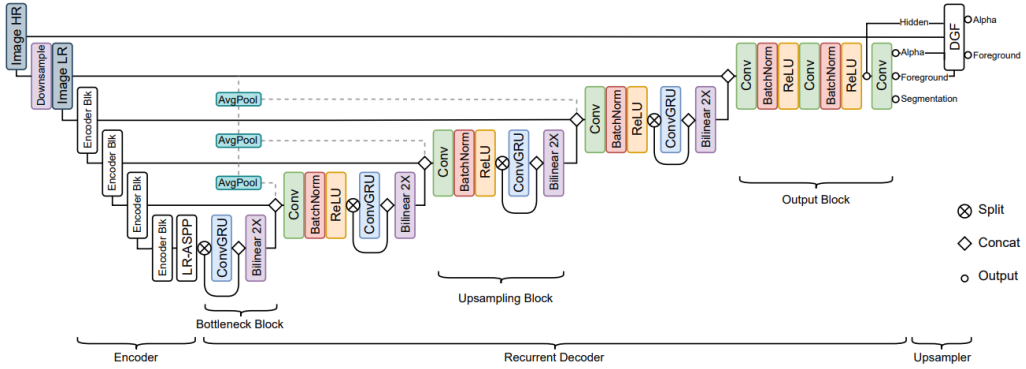
相对于**腾讯NCNN**等工业界开源的最佳实现，**加速3倍以上**

CVPR 2020 (CCF A会议) IJCV 2023 (CCF A期刊)

卷积结构量化：准确高效表征量化结构 (视频抠图)

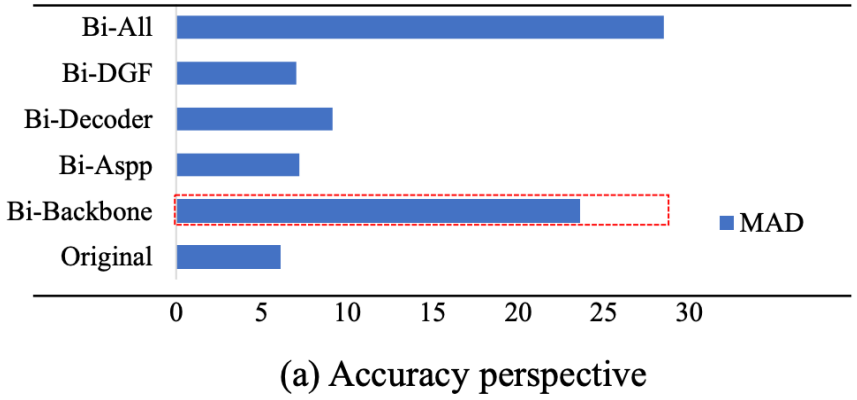
问题： 编码与解码器二值量化后信息提取与利用效率差

思路： 打破原始结构在量化背景下的精度和效率瓶颈

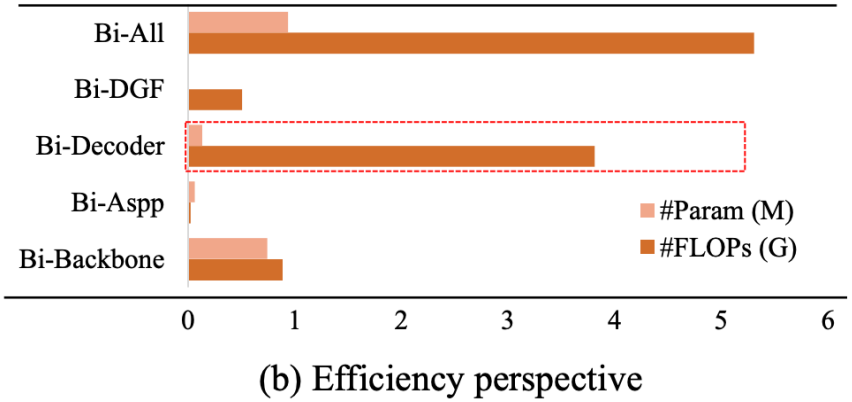


$$I = \alpha F + (1 - \alpha)B.$$

二值量化抠图模型的准确率和效率瓶颈



(a) Accuracy perspective



(b) Efficiency perspective

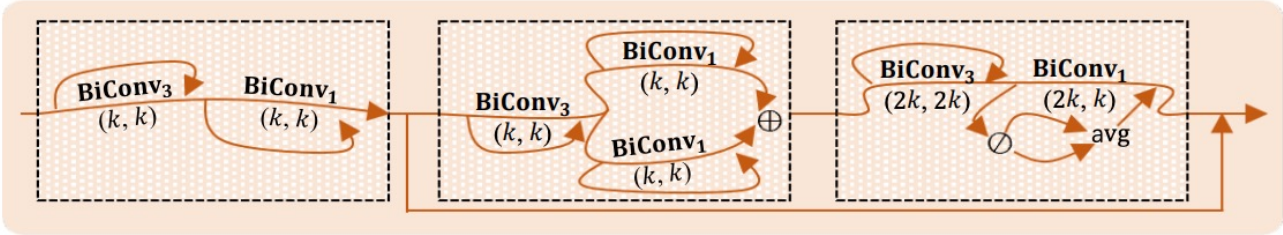
NeurIPS 2023 (CCF A会议)

卷积结构量化：准确高效表征量化结构 (视频抠图)

问题： 编码与解码器二值量化后信息提取与利用效率差

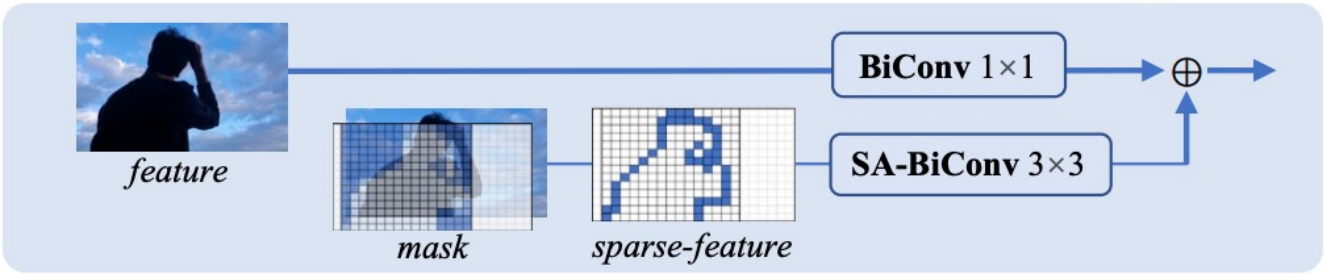
思路： 打破原始结构在量化背景下的精度和效率瓶颈

准确编码：
密集弹性连接构建
准确二值骨干



Shrinkable Binarized Block (SBB)

$$\textbf{\textit{SBB}}: \quad \boldsymbol{o} = \theta^{\text{dn}} \cdot \theta^{\text{up}}(\boldsymbol{x}') + \boldsymbol{x}', \quad \boldsymbol{x}' = \theta^{\text{eq}}(\boldsymbol{x})[c^{\boldsymbol{x}} = c^{\boldsymbol{o}}] + \theta^{\text{up}}(\boldsymbol{x})\left[c^{\boldsymbol{x}} = \frac{1}{2}c^{\boldsymbol{o}}\right].$$



Sparse-Assisted Binarization (SAB)

$$\textbf{\textit{SAB}}: \quad \boldsymbol{o} = \text{SA-BiConv}_3(\boldsymbol{x}; \text{bilinear}^k(M_{\text{inc}})) + \text{BiConv}_1(\boldsymbol{x}),$$

高效解码：
稀疏遮罩计算实现
高效表征利用

NeurIPS 2023 (CCF A会议)

卷积结构量化：准确高效表征量化结构 (视频抠图)

极限压缩后媲美工业级视频抠图模型 原始性能

12倍加速　　24倍压缩　　全精度级别性能

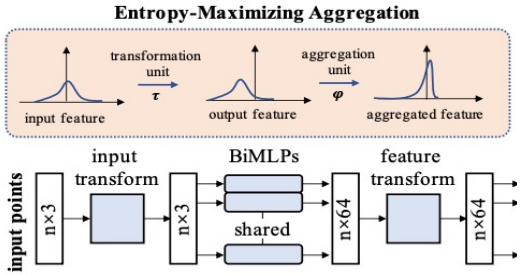
NeurIPS 2023 (CCF A会议)

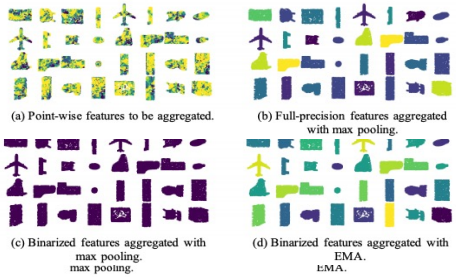
全连接结构量化：面向量化结构表征异质化（点云分类、分割）

问题： 如何减小二值化对线性单元和其相关结构的影响

思路： 聚合时防止同化特征，面向结构恢复特征分布

BiPointNet: 首次面向3D视觉任务开展量化研究

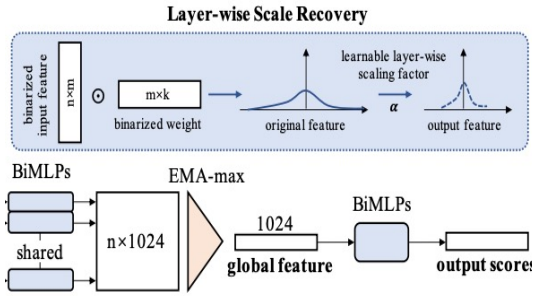


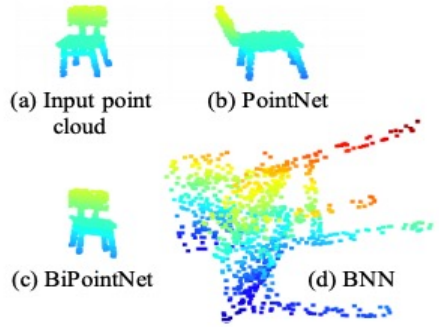


特征平移/异质聚合

$$\mathbf{Y} = \mathrm{EMA}(\mathbf{X}_\phi) = \varphi(\tau(\mathbf{X}_\phi))$$

优化大尺寸聚合操作： 提出熵最大化聚合函数，解决特征聚合同质化问题





尺度放缩/方差归一

$$\alpha_0 = \frac{\sigma(\mathbf{A} \otimes \mathbf{W})}{\sigma(\mathbf{B}_a \odot \mathbf{B}_w)}$$

恢复扭曲尺度的特征： 提出逐层尺度恢复因子，解决特征尺度扭曲问题

ICLR 2021 (机器学习顶会)

全连接结构量化：面向量化结构表征异质化 (点云分类、分割)

3D点云分类等多项任务实际部署 存储最省

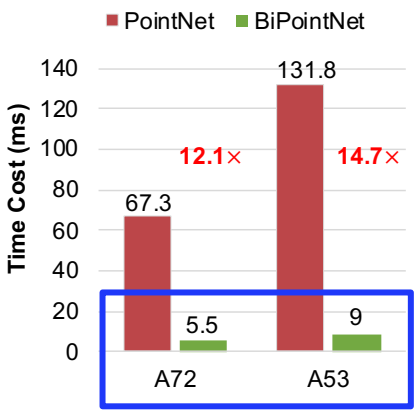
方法	机构	精度	文献
BiPointNet	**北京航空航天大学**	**86.40%**	**ICLR 2021**
XNOR-Net	华盛顿大学	64.90%	NeurIPS 2016
ABC-Net	大疆科技	4.10%	NeurIPS 2017
BiReal-Net	香港科技大学, 腾讯	4.00%	ECCV 2020
XNOR++	三星	4.10%	BMVC 2019
BNN	蒙特利尔大学	7.10%	NeurIPS 2015

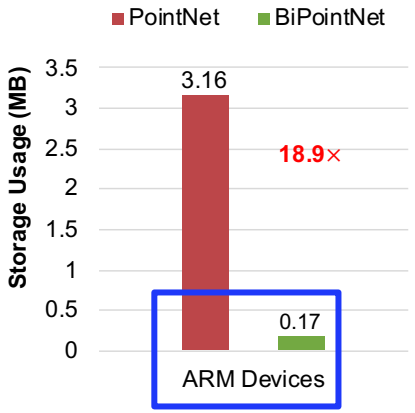
二值模型精度比较 (PointNet, ModelNet40)

北京地区广受关注学术论文
(国内图形图像领域13项)

被邀引入**商汤OpenMM**，**亚马逊DGL**等**知名开源平台**

Star 12,030





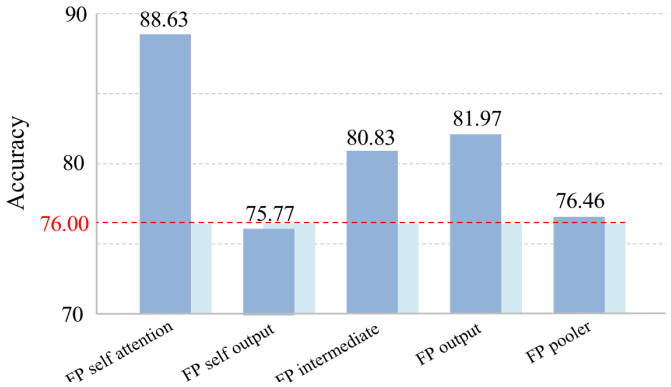
在**真实边缘设备**上部署并实现
14.7倍加速和**18.9倍压缩**

ICLR 2021 (机器学习顶会)

Transformer 量化训练：二值注意力机制恢复

问题：如何使基于二值参数的注意力机制发挥作用

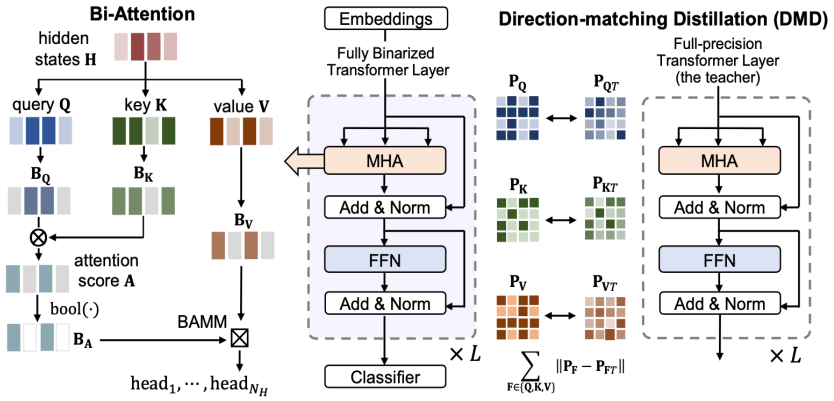
思路：恢复二值注意力机制，对齐注意力量化优化方向



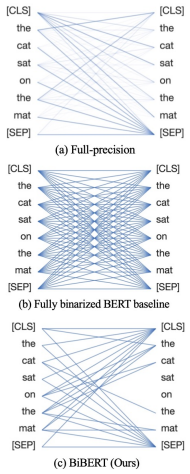
注意力模块的量化对模型精度起到关键影响

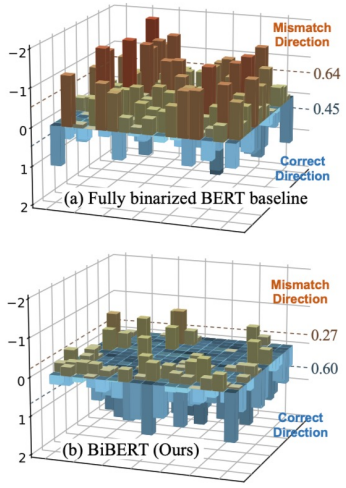
关注于注意力模块的量化与优化方法

(1) 提出二值注意力，恢复注意力机制作用
(2) 方向对齐蒸馏框架，优化模型收敛效果



二值友好的Transformer蒸馏量化框架





ICLR 2022 (机器学习顶会)

Transformer 量化训练：二值注意力机制恢复

同时期Transformer量化工作中二值化计算 转化最全

方法	机构	精度	计算	文献
BiBERT	**北京航空航天大学**	**61.0%**	**0.4**	**ICLR 2022**
TernaryBERT (2-2-2)	华为诺亚	45.5%	1.5	EMNLP 2020
TernaryBERT (2-2-1)	华为诺亚	42.3%	0.8	EMNLP 2020
BinaryBERT (1-1-2)	华为诺亚	53.7%	0.8	ACL 2021
BinaryBERT(1-1-1)	华为诺亚	41.0%	0.4	ACL 2021
Q2BERT	英特尔	47.7%	6.5	NeurIPS 2019

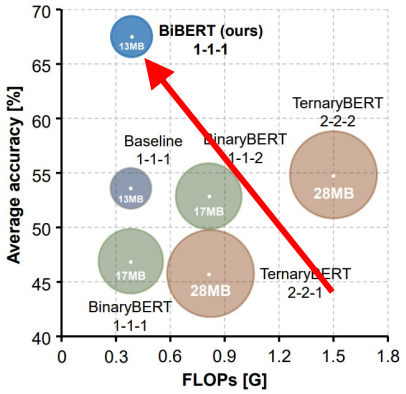
二值模型精度比较 (BERT, GLUE Benchmark)

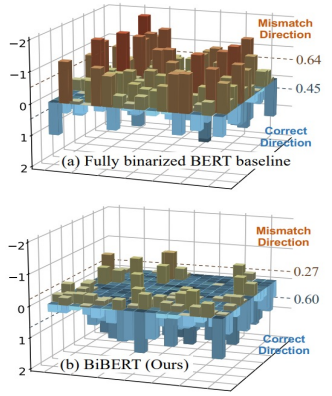
国际上首次实现BERT全二值化
国际上首提二值注意力硬件实现

CCF-松果研究基金（全球**36**项）

算法被**知名深度学习开源平台 PaddlePaddle** 收录

Star 1,423



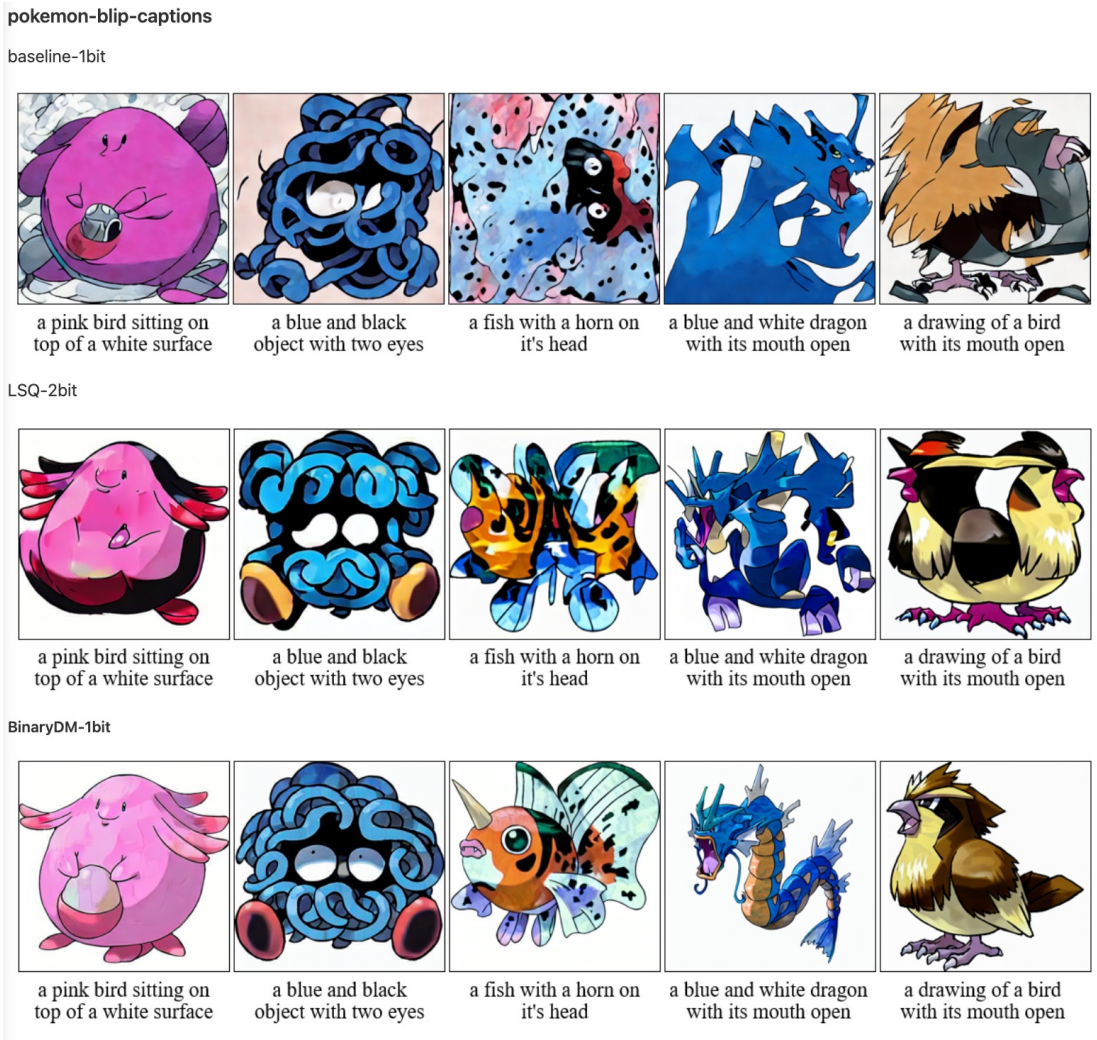


精度远超相同甚至**更高位宽**下的现有Transformer量化工作

推理最大程度转为按位运算**31.2倍存储**、**51.3倍计算**理论节省

ICLR 2022 (机器学习顶会)

Transformer 量化训练：二值注意力机制恢复 (生成任务)



二值量化DM基线

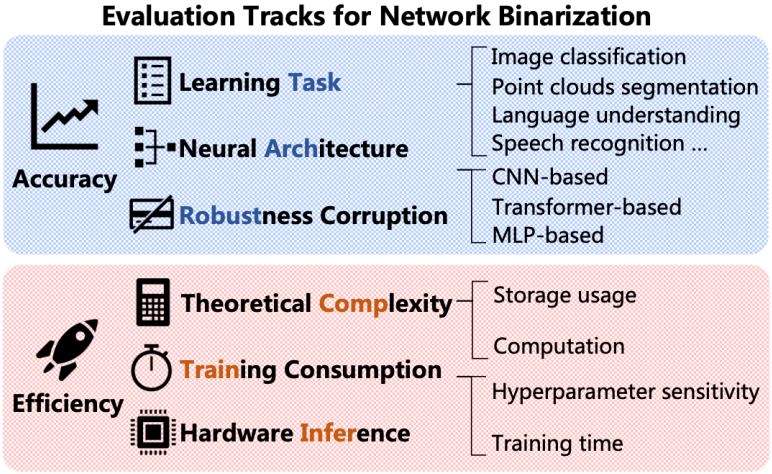
2-bit SOTA量化

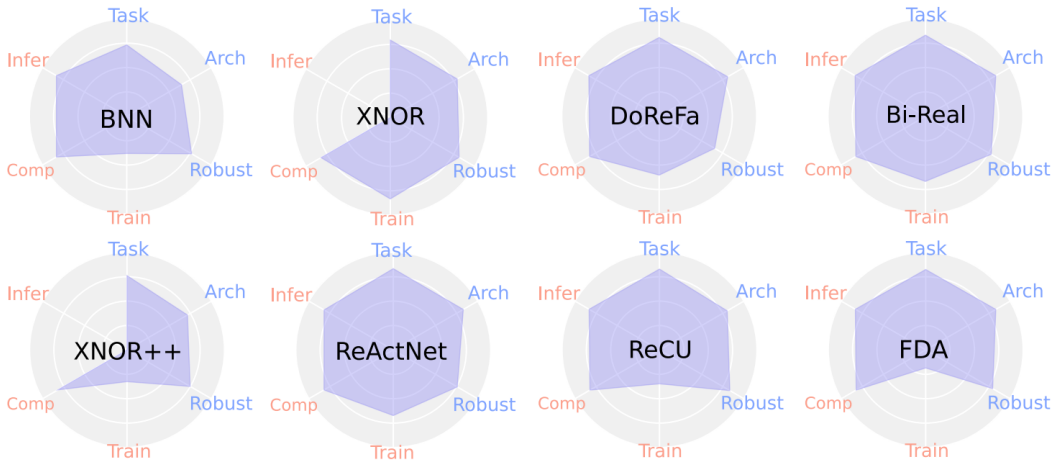
我们的 1-bit 量化

Submission 2024

综述与基准

基准：首个二值量化基准与分析 BiBench





6 评估轨道: 准确率、效率

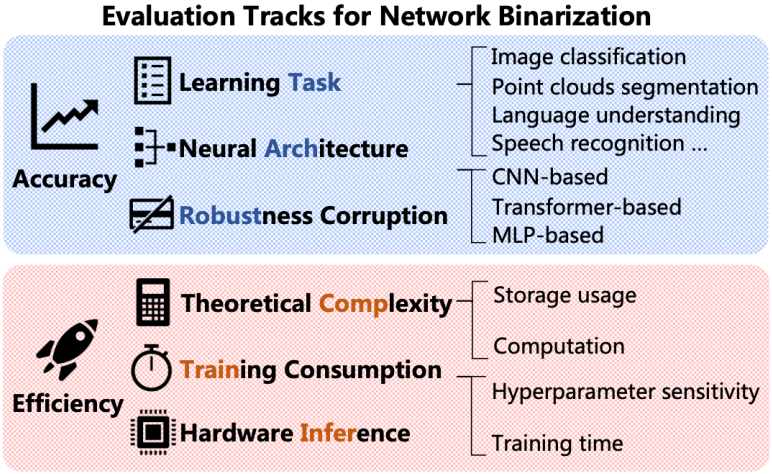
4 GPU年（Nvidia V100）

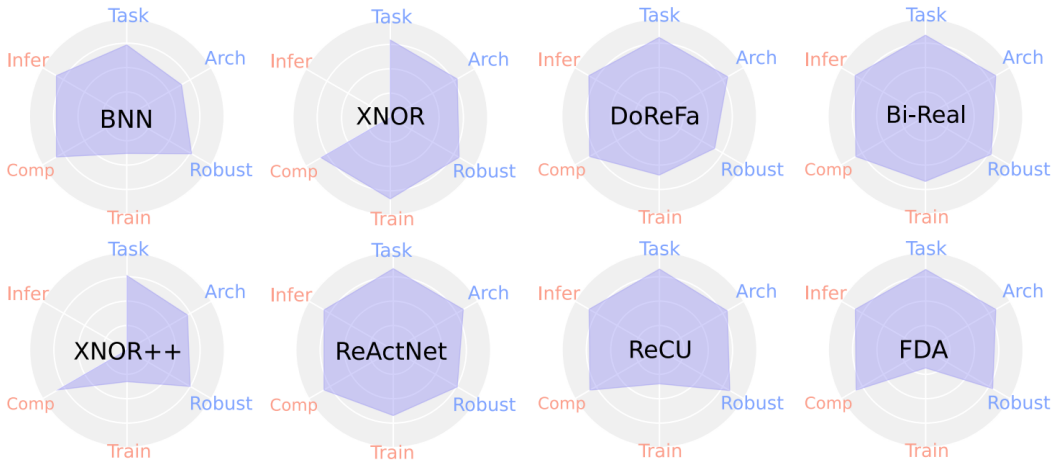
8 二值量化算法
9 深度学习数据集
（5 个2D/3D视觉数据集）
13 神经网络架构
2 硬件部署框架
14 边缘硬件芯片

ICML 2023 (CCF A会议)

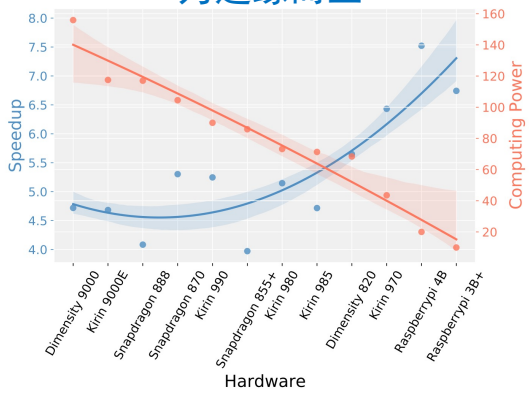
综述与基准

基准：首个二值量化基准与分析 BiBench





为边缘而生



开展了二值量化领域的**首次面向软硬件的大规模全面评测**

揭示了二值量化**软硬件特性**，以及**准确高效的算子设计范式**

ICML 2023 (CCF A会议)

综述与基准

Pattern Recognition

Available online 21 February 2020, 107281

In Press, Corrected Proof ⓘ

Binary neural networks: A survey

Haotong Qin [a], Ruihao Gong [a], Xianglong Liu [a, b] ✉, Xiao Bai [e], Jingkuan Song [c], Nicu Sebe [d]

⊞ Show more

https://doi.org/10.1016/j.patcog.2020.107281

Get rights and content

中文版：二值神经网络（Binary Neural Networks）综述

PaperWeekly: https://mp.weixin.qq.com/s/QGva6fow9tad_daZ_G2p0Q

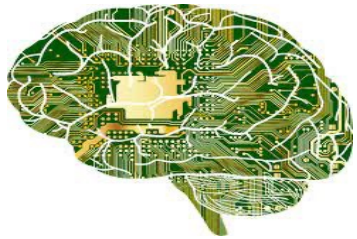
Pattern Recognition 期刊 “**Most Cited Papers 2023**”

ESI 高被引论文 2022

UC Berkely 2022年量化最新综述《A Survey of Quantization Methods for Efficient Neural Network Inference》沿用了我们文章中对二值网络方法的分类

Pattern Recognition 2020 (CCF B期刊)

未来展望



现实需求	AIoT设备上的 广泛视觉模型需求	预训练视觉生成 模型的流行	新兴计算和硬件形 式不断涌现
研究展望	边缘侧视觉模型的 训练、压缩和部署	有限资源下的视觉 生成模型压缩方法	面向事件相机的脉 冲神经网络等

边缘机器学习 *生成模型压缩* *新型智能计算*

中国图象图形学学会第九期学生会员分享论坛

感谢各位专家老师和同学们！

秦浩桐
ETH Zurich
2024/03/31