哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY

PD-Quant: Post-Training Quantization based on Prediction Difference Metric

Jiawei Liu[1,⋆,⋄] Lin Niu[1,⋆,⋄] Zhihang Yuan[2,†] Dawei Yang[2] Xinggang Wang[1] Wenyu Liu[1,†]
[1] School of EIC, Huazhong University of Science & Technology [2] Houmo AI
{jiaweiliu, linniu}@hust.edu.cn hahnyuan@gmail.com dawei.yang@houmo.ai
{xgwang, liuwy}@hust.edu.cn

Chao Zeng
2023/10/31

目录页

CONTENTS

1 Motivation

- 当前PTQ量化中对scale的选择主要使用MSE、cosine distance等局部信息进行优化，未考虑到模型量化的全局误差。
- PTQ量化中校验集往往较小，容易存在过拟合问题

2 Method

Prediction Difference Loss

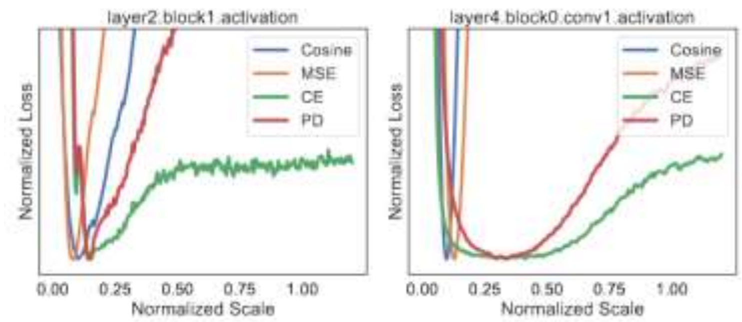
MSE、Cosine distance等指标不能不很好的接近真实的 task loss

$$\arg\min_{S_a} \mathcal{L}_{PD}(f_l(\tilde{A}_{l-1}),\ f_{l+1}(L_l^q(\tilde{A}_{l-1}))), \qquad (3)$$

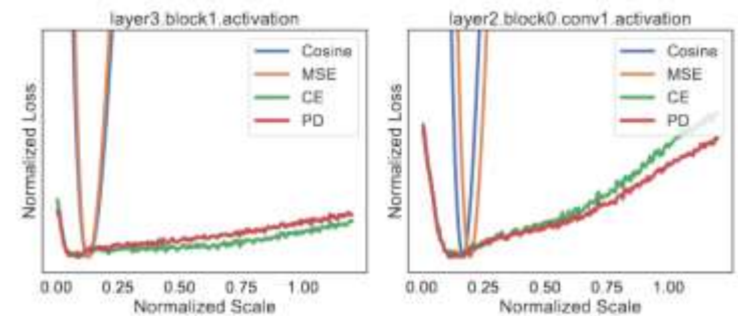
将量化的激活$\tilde{A}_l$ 送入FP层获得输出的预测值

损失函数计算输出的KL散度作为损失度量

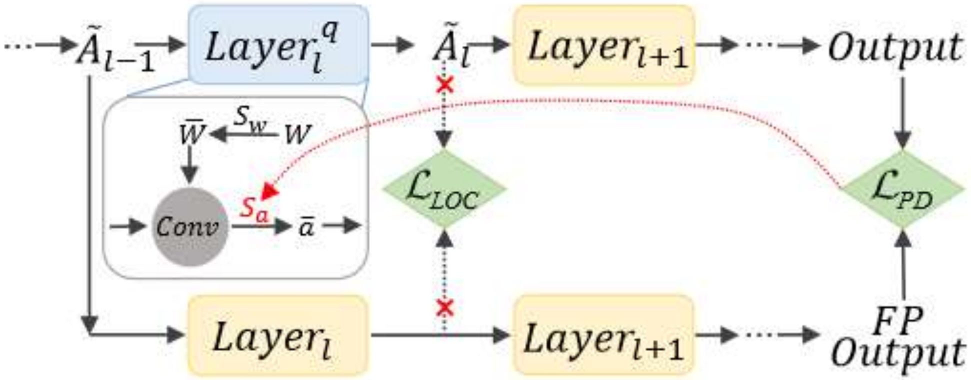
S_a 的表示激活的scale factor



(a) ResNet-18



(b) ResNet-50



PD Loss 计算图解

Model	ResNet-18		ResNet-50		RegNet-600M	
Bits	W8A2	W4A2	W8A2	W4A2	W8A2	W4A2
Min-Max^l	-	-	-	-	-	-
Cosine^l	11.09	4.15	2.19	1.14	0.96	0.65
MSE^l	23.15	10.31	9.23	4.85	3.71	1.88
PD^g	**28.41**	**12.27**	**11.31**	**6.01**	**7.47**	**3.17**

PD Loss使用的全局损失相比与局部损失拥有更好的优化效果，极大的提高模型性能。

PD Loss主要考虑了Scale的求解优化，AdaRound和BRECQ同时还通过如下考虑了Round误差

$$\widetilde{\mathbf{W}} = \mathrm{s} \cdot clip\left(\left\lfloor \frac{\mathbf{W}}{\mathrm{s}} \right\rfloor + h\left(\mathbf{V}\right), \mathrm{n}, \mathrm{p}\right).$$

AdaRound

$$\hat{\mathbf{w}} = s \times \mathrm{clip}\left(\lfloor \frac{\mathbf{w}}{s} \rfloor + \sigma(\mathbf{v}), n, p\right)$$

BRECQ

PQ-Quant中通过优化变量θ来决定在舍入时向上舍入还是向下舍入

$$\tilde{x} = clamp\left(\lfloor \frac{x+\theta}{S} \rceil + Z;\ q_{min}, q_{max}\right), \qquad (4)$$

Method	Bits (W/A)	Acc(val)	Acc(cali)
FP	32/32	71.01	70.90
PD-only	2/2	1.07	70.51
PD+Reg		**49.16**	71.09
PD+Reg+Drop		**52.74**	68.26
PD-only	4/2	51.32	70.41
PD+Reg		**56.20**	70.41
PD+Reg+Drop		**58.17**	68.36

PD-only的时候校验集过拟合

B^q表示量化函数用于对第l块进行量化

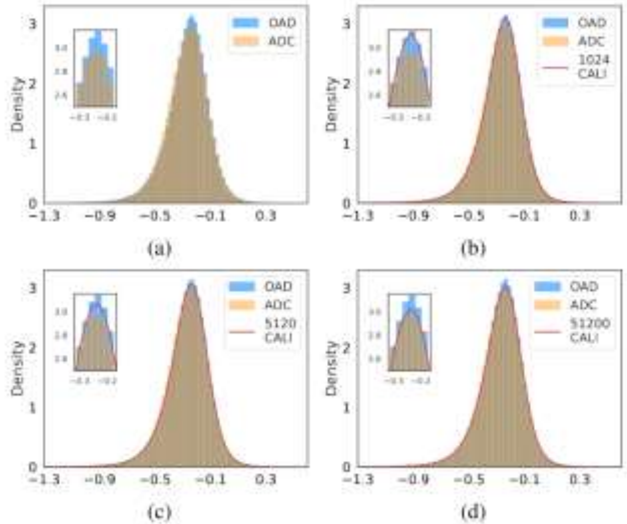
L_{reg}通过计算量化前后输出的MSE值缓解过拟合问题

$$\arg\min_{\theta,S_a} \mathcal{L}_{PD}(O_{fp},\ f_{l+1}(\tilde{A}_l) + \lambda_r \mathcal{L}_{reg}(A_l, \tilde{A}_l)), \quad (5)$$
$$\tilde{A}_l = B_l^q(\tilde{A}_{l-1};\ \theta, S_a),$$

CNN网络中存储了原始训练数据集的均值和方差，能更好的反映整体数据分布

DC的目的就是让校验数据集的分布尽量满足原始训练数据的分布，减少模型过拟合

$$\arg\min_{A_{l-1}^{DC}} \lambda_c \sum_{i=1}^{n} (\| \hat{\mu}_{(i,l)} - \mu_{(i,l)} \|_2^2 + \| \hat{\sigma}_{(i,l)} - \sigma_{(i,l)} \|_2^2) + \| A_{l-1}^{DC} - A_{l-1}^{FP} \|_2^2, \quad (6)$$



当校验集增大时候，校验集的和密度曲线接近DC校验后的分布，ADC的分布更接近数据的真实分布

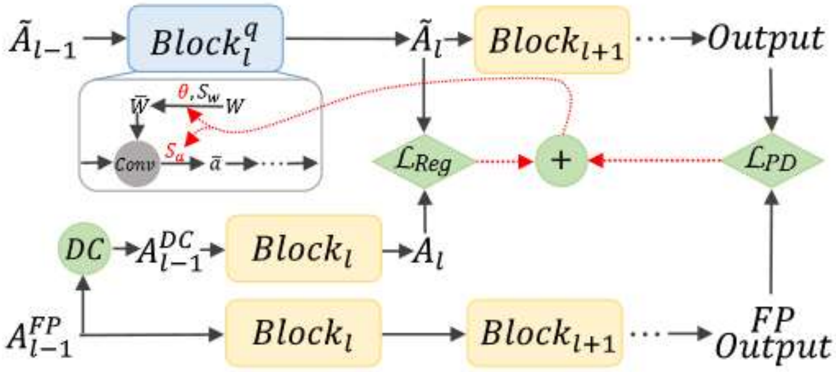


Figure 3. An overview of the PD-Quant. The blue and yellow rectangles indicate the quantized and FP layer, respectively. The green diamond is marked as the loss function. The green circle with DC indicates the Distribution Correction. FP output is the prediction of the whole FP network.

PD-Quant整体框架图

3 Evaluation

基础试验设置

评估模型：ResNet、MobileNetV2、RegNet、MNasNet

校验集大小：1024来自ImageNet

Batch Size: 32

Activation quantization 学习率：4e-5

Weight quantization Rounding学习率：3e-3

DC 学习率：1e-3

PD-Quant在低比特位下效果存在明显提升

Methods	Bits (W/A)	ResNet-18	ResNet-50	MobileNetV2	RegNetX-600MF	RegNetX-3.2GF	MNasx2
Full Prec.	32/32	71.01	76.63	72.62	73.52	78.46	76.52
ACIQ-Mix [1]	4/4	67.00	73.80	-	-	-	-
LAPQ [33]	4/4	60.30	70.00	49.70	57.71	55.89	65.32
Bit-Split [46]	4/4	67.56	73.71	-	-	-	-
AdaRound [31]	4/4	67.96	73.88	61.52	68.20	73.85	68.86
QDrop [47]*	4/4	69.17	75.15	68.07	70.91	76.40	72.81
PD-Quant	4/4	**69.23**±0.06	**75.16**±0.07	**68.19**±0.12	**70.95**±0.12	**76.65**±0.09	**73.26**±0.09
LAPQ	2/4	0.18	0.14	0.13	0.17	0.12	0.18
Adaround	2/4	0.11	0.12	0.15	-	-	-
QDrop*	2/4	64.57	70.09	53.37	63.18	71.96	63.23
PD-Quant	2/4	**65.17**±0.08	**70.77**±0.15	**55.17**±0.28	**63.89**±0.13	**72.38**±0.11	**63.40**±0.21
QDrop*	4/2	57.56	63.26	17.30	49.73	62.00	34.12
PD-Quant	4/2	**58.59**±0.15	**64.18**±0.14	**20.10**±0.37	**51.09**±0.15	**62.79**±0.13	**39.13**±0.51
QDrop*	2/2	51.42	55.45	10.28	39.01	54.38	23.59
PD-Quant	2/2	**53.14**±0.14	**57.16**±0.15	**13.76**±0.40	**40.67**±0.26	**55.06**±0.23	**27.58**±0.60

Table 3. Comparison on PD-Quant with various post-training quantization algorithms. * denotes our implementation using open-source codes. PD-Quant is our proposed method. Other results listed are all from [47]. We gain the results of 10 runs using randomly sampled calibration sets. The results in the table include the mean and standard deviation.

Model	ResNet-18		MobileNetV2	
Bits	W2A2	W4A2	W2A2	W4A2
QDrop	51.42	57.56	10.28	17.30
PD-only	1.07	51.32	7.01	13.59
PD+Reg	52.74	58.17	13.49	20.05
QDrop+DC	52.32	57.77	10.38	17.58
PD-Quant	**53.08**	**58.65**	**14.17**	**20.40**

Table 5. Ablation study (top-1 accuracy(%)) on validation set for our proposed method. QDrop is the baseline method. PD-only means optimizing quantization parameters by only PD loss. Reg means regularization. PD-Quant is our proposed method, including PD, Reg, and DC for optimizing both activation scaling factors and rounding values.

Method	ResNet-18	MobileNetV2	RegNetX-600MF
QDrop	0.43h	0.93h	0.89h
PD	0.91h	2.26h	2.37h
PD+DC	1.11h	2.68h	2.75h

Table 10. Time cost comparison. (one Nvidia RTX A6000)

PD-Quant需要时间进行少量的微调

Fine-tuning 20000 iterations

4 Summary

- 实现了activation的2bit量化
- 与QDrop结合实现了W2A2
- 对scale选择时考虑了全局信息
- 通过对激活分布调整缓解了校验的过拟合问题

哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY

恳请各位老师批评指正