神经网络低比特量化——LSQ+

发布于 2020-08-11 02:39:10 2.2K 0 举报

文章被收录于专栏：AI异构

LSQ+

> 本文为高通基于LSQ改进的量化工作，论文题目为：LSQ+: Improving low-bit quantization through learnable offsets and better initialization。

摘要

与ReLU不同，在流行的高效架构中经常使用的较新的激活函数（如Swish，H-swish，Mish）也可能导致负激活值，正负范围出现偏差。典型的可学习量化方案（PACT,LSQ）假设激活无符号量化，并将所有负激活量化为零，这会导致性能显着下降。简单的使用带符号的量化来容纳这些负值需要一个额外的符号位，这对于低位（2位，3位，4位）量化而言是昂贵的。为了解决这个问题，我们提出了LSQ 的扩展，即LSQ +，该方法引入了一种通用的**非对称量化**方案，该方案具有**可训练的尺度和偏移参数，可以学会适应负激活**。基于梯度的可学习量化方案通常会有最终训练性能高度不稳定性的缺点，因此需要大量的超参数调整才能达到令人满意的性能。LSQ+ 通过**对量化参数使用基于 MSE 的初始化方案**来缓解此问题。本文表明，这种初始化使得在多次训练中最终表现的差异明显降低。总体而言，LSQ+ 显示了 EfficientNet 和 MixNet 的最新结果，并且在具有 Swish 激活的神经网络的低位量化方面也明显优于 LSQ（例如：在ImageNet数据集上，对EfficientNet-B0进行W4A4量化处理，获得了1.8%的增益；对EfficientNet-B0进行W2A2量化处理，获得了5.6%的增益）。

方法

Learnable asymmetric quantization

量化公式

$$\bar{x} = \left\lfloor \operatorname{clamp}\left(\frac{x-\beta}{s}, n, p\right)\right\rceil$$
$$\hat{x} = \bar{x} \times s + \beta$$

量化反向求导

$$\frac{\partial \hat{x}}{\partial s} = \frac{\partial \bar{x}}{\partial s} s + \bar{x} \simeq \begin{cases} -\frac{x-\beta}{s} + \left\lfloor \frac{x-\beta}{s} \right\rceil & \text{if } n < \frac{x-\beta}{s} < p \\ n \text{ or } p & \text{otherwise.} \end{cases}$$

$$\frac{\partial \hat{x}}{\partial \beta} = \frac{\partial \bar{x}}{\partial \beta} s + 1 \simeq \begin{cases} 0 & \text{if } n < (x-\beta)/s < p \\ 1 & \text{otherwise.} \end{cases}$$

参数与激活量化形式

对于权重量化，使用对称符号量化，因为可以凭经验观察到层权重在零附近对称分布。激活采用非对称量化形式。与对称量化相比，激活的非对称量化在推理过程中没有额外的成本，因为可以在编译时预先计算附加的偏移项并将其合并到偏差中：

$$\hat{w}\hat{x} = (\bar{w} \times s_w)(\bar{x} \times s_x + \beta) = \bar{w}\bar{x}s_w s_x + \underbrace{\beta s_w \bar{w}}_{\text{bias}}$$

LSQ+四种配置方案

Table 1. Different possible parametrizations for LSQ+'s learnable asymmetric quantization scheme

Configuration	s	β	n	p
Config 1 : Unsigned + Symmetric (LSQ)	trainable	N/A	0	2^b-1
Config 2 : Signed + Symmetric	trainable	N/A	-2^{b-1}	$2^{b-1}-1$
Config 3 : Signed + Asymmetric	trainable	trainable	-2^{b-1}	$2^{b-1}-1$
Config 4 : Unsigned + Asymmetric	trainable	trainable	0	2^b-1

目录

LSQ+不同配置方案

参数初始化方法

Scale initialization for weight quantization

$$s_{\text{init}} = \max(|\mu - 3 * \sigma|, |\mu + 3 * \sigma|)/2^{b-1}$$

$$\mu$$

为权重绝对值的均值

$$\sigma$$

为权重绝对值的方差

Scale/offset initialization for activation quantization

通过优化MSE最小化问题来初始化每层的比例和偏移参数

$$s_{\text{init}}, \beta_{\text{init}} = \underset{s,\beta}{\arg\min} ||\hat{x} - x||_F^2$$

将 s 和 β 的梯度嵌入到 PyTorch 的自动求导功能中，以便在几批数据中优化{

$$s_{\text{init}}, \beta_{\text{init}}$$

}。

实验结果

EfficientNet-B0

Table 2. Comparison of all configurations of quantization with EfficientNet-B0 (FP accuracy: 76.1%)

Method	W2A2	W3A3	W4A4
Config 1 : LSQ (Unsigned + Symmetric)	43.5%	67.5%	71.9%
Config 2 : Signed + Symmetric	23.7%	54.8%	68.8%
Config 3 : Signed + Asymmetric	**49.1%**	**69.9%**	73.5%
Config 4 : Unsigned + Asymmetric	48.7%	69.3%	**73.8%**

EfficientNet-B0不同配置下的结果对比

MixNet-S

Table 3. Comparison of all configurations of quantization with MixNet-S (FP accuracy: 75.9%)

Method	W2A2	W3A3	W4A4
Config 1 : LSQ (Unsigned + Symmetric)	39.9%	64.3%	70.4%
Config 2 : Signed + Symmetric	23.4%	62.1%	67.2%
Config 3 : Signed + Asymmetric	42.5%	**66.7%**	71.6%
Config 4 : Unsigned + Asymmetric	**42.8%**	66.1%	**71.7%**

MixNet-S不同配置下的结果对比

ResNet18

Table 4. Comparison of all configurations of quantization with ResNet18 (FP accuracy: 70.1%)

Method	W2A2	W3A3	W4A4
PACT [5]	64.4%	68.1%	69.2%
DSQ [8]	65.2%	68.7%	69.6%
QIL [14]	65.7%	69.2%	70.1%
Config 1 : LSQ (Unsigned + Symmetric)	66.7%	**69.4%**	70.7%
Config 2 : Signed + Symmetric	64.7%	66.1%	69.2%
Config 3 : Signed + Asymmetric	66.7%	**69.4%**	70.7%
Config 4 : Unsigned + Asymmetric	**66.8%**	69.3%	**70.8%**

ResNet18不同配置下的结果对比

领券

消融实验

初始化结果稳定性分析

Table 5. Δ_{acc} around mean accuracy across 5 training runs for EfficientNet quantization using Config 4 with different initializations. Note: other tables show the *best* accuracy after grid search on hyperparameters, which is different from mean accuracy.

Quantization Parameter Initialization	Mean Acc $\pm\Delta_{acc}$	
	W4A4	W2A2
Mix-max	$71.3 \pm 2.2\%$	$43.8 \pm 4.7\%$
LSQ	$72.0 \pm 1.6\%$	$44.4 \pm 2.9\%$
LSQ+	$73.0 \pm \mathbf{0.9\%}$	$46.8 \pm \mathbf{1.9\%}$

初始化结果稳定性分析

非对称偏置的作用

Table 6. Performance difference between fixed and learned offset for EfficientNet quantization at W4A4 using Config 4

Method	W4A4
Fixed $\beta = 0$ (LSQ)	71.9%
Fixed $\beta = x_{min}$	72.5%
Learned β	73.8%

非对称偏置的作用

更多内容关注微信公众号【AI异构】