

Quantization Error Propagation: Revisiting Layer-Wise Post-Training Quantization

1.研究背景

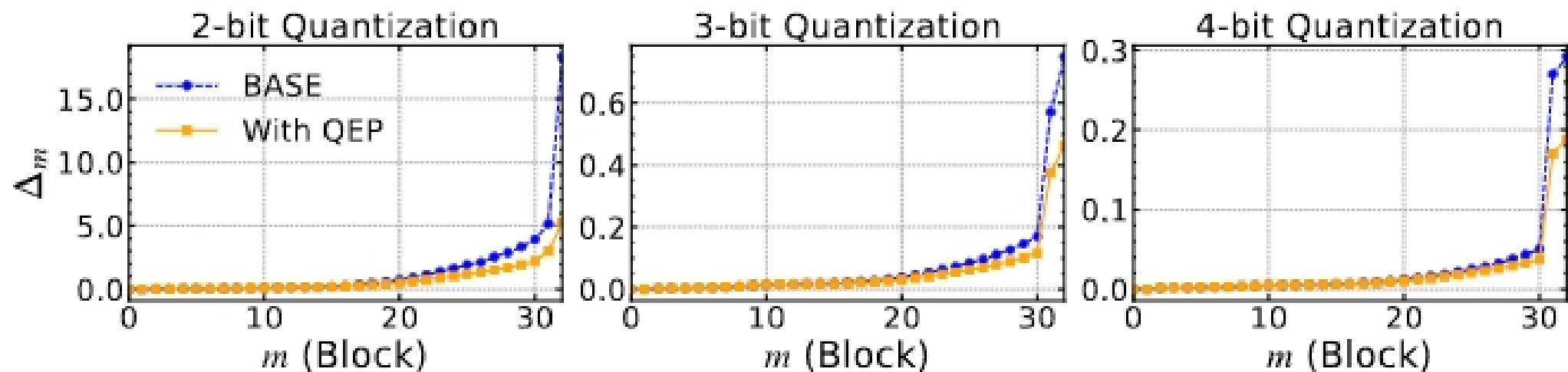
- 大模型部署痛点：参数量大导致显存与推理开销高，后训练量化是主流方案
- 分层PTQ的核心优势：无需重新训练、逐层独立量化、算力内存需求低，应用最广泛
- 代表方法：RTN、GPTQ、AWQ、QuIP
- 现有方法共性盲区：单层独立优化权重，完全忽略量化误差跨层传递与累积
- 核心问题：量化误差逐层放大

2.核心瓶颈： 量化误差逐层累计

验证实验设计

模型：Llama2-7B，仅量化前10个Transformer Block，后续层保持全精度

指标：逐层计算全精度与量化模型输出的Frobenius范数误差



3.QEP核心方法： 重构优化目标

传统PTQ优化目标：
$$\min_{\widehat{\mathbf{W}}_l \in \mathbb{Q}^{n_l \times d_l}} \left\| \mathbf{W}_l \mathbf{X}_l - \widehat{\mathbf{W}}_l \mathbf{X}_l \right\|_F^2$$

QEP的优化目标：
$$\min_{\widehat{\mathbf{W}}_l \in \mathbb{Q}^{n_l \times d_l}} \left\| \mathbf{W}_l \mathbf{X}_l - \widehat{\mathbf{W}}_l \widehat{\mathbf{X}}_l \right\|_F^2$$

核心思路：当前层量化时，考虑上游累积的量化误差，主动对权重做补偿矫正

缺点：破坏了原有PTQ框架中实现高效量化的关键结构，使得量化无法使用现有基于海森矩阵的加速

3.QEP核心方法：权重矫正与可控传播

(1).权重矫正

$$\mathbf{W}_l^* := \mathbf{W}_l + \mathbf{W}_l \delta_l \widehat{\mathbf{X}}_l^\top \widehat{\mathbf{H}}_l^{-1} = \underset{\widehat{\mathbf{W}}_l \in \mathbb{R}^{n_l \times d_l}}{\operatorname{argmin}} \left\| \mathbf{W}_l \mathbf{X}_l - \widehat{\mathbf{W}}_l \widehat{\mathbf{X}}_l \right\|_F^2$$

where $\delta_l := \mathbf{X}_l - \widehat{\mathbf{X}}_l$ represents the accumulated quantization error from proceeding layers, $\widehat{\mathbf{H}}_l := \widehat{\mathbf{X}}_l \widehat{\mathbf{X}}_l^\top$ denotes the empirical Hessian constructed from quantized activations.

$$\min_{\widehat{\mathbf{W}}_l \in \mathbb{Q}^{n_l \times d_l}} \left\| \mathbf{W}_l^* \widehat{\mathbf{X}}_l - \widehat{\mathbf{W}}_l \widehat{\mathbf{X}}_l \right\|_F^2$$

3.QEP核心方法：权重矫正与可控传播

(2).可控传播

当校准数据集规模较小且无法充分代表目标任务，或者模型包含参数量较多的模块时容易出现过拟合，为解决这一问题需对矫正项进行缩放

$$\mathbf{W}_l^*(\alpha_l) = \mathbf{W}_l + \alpha_l \mathbf{W}_l \delta_l \widehat{\mathbf{X}}_l^\top \widehat{\mathbf{H}}_l^{-1}$$

若设置缩放值=1对应完全矫正的情况，即QEP的优化目标

若假设 $\mathbf{x}_l = \bar{\mathbf{x}}_l$ 且缩放值=0对应传统优化目标

4.实验

基准方法：RTN、GPTQ、AWQ、QuIP

精度：INT4、INT3、INT2

α ：Llama-7 70B模型的MLP层为0(多层感知器层会带来高昂的矫正项计算成本)，其余模型所有层均为0.5

数据集：GPTQ与QuIP使用C4数据集，AWQ使用Pile数据集

模型：Llama-2和Llama-3系列，参数规模从7B至70B，以及Mistral-7B

评估：采用WikiText2数据集、PTB数据集和C4数据集的困惑度以及ArcE基准、PiQA基准和SC基准的零样本准确率对量化模型进行评估

硬件：NVIDIA V100 GPU