

模型压缩

《大语言模型》编写团队：唐天一

➤ 量化：将映射浮点数到整数的过程

$$X_q = R(X/S) - Z$$

➤ 反量化：从量化值中恢复原始值

$$\tilde{X} = S \cdot (X_q + Z)$$

➤ 量化误差：原始值 X 和恢复值 \tilde{X} 之间的数值差异

$$\Delta = \|X - \tilde{X}\|_2^2$$

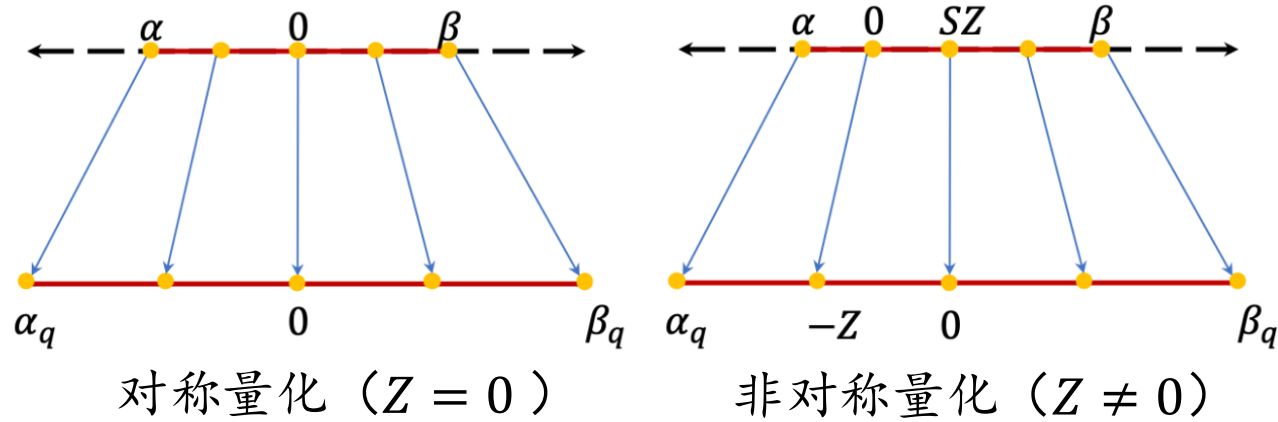
R：取整函数

S：放缩因子

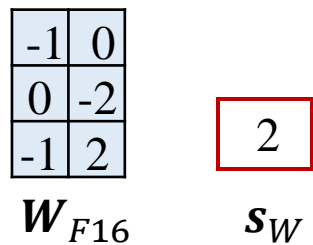
Z：零点因子

量化的目标是
最小化量化误差

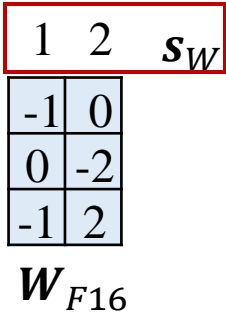
- 对称量化和非对称量化
 - 根据零点因子 Z 是否为零



- 量化粒度的选择



张量量化：一个矩阵
定义一组量化参数



通道量化：一个矩阵对列维度
设置特定的量化参数

➤ 非对称量化方法计算示例

$$\mathbf{X} = [[1.2, 2.4, 3.6], [11.2, 12.4, 13.6]]$$

确定输入范围 $\in [1.2, 13.6]$

$$S \cdot (127 + Z) = 13.6$$

量化到整数范围 $[-128, 127]$
两个边界值映射

$$S \cdot (-128 + Z) = 1.2$$

$$\Rightarrow S = 0.0486, \quad Z = -152$$

计算量化参数 S 和 Z

$$\mathbf{X}_q = [[-127, -103, -78], [78, 103, 127]]$$

计算量化后结果 \mathbf{X}_q

$$\tilde{\mathbf{X}} = [[1.22, 2.38, 3.60], [11.18, 12.40, 13.58]]$$

计算反量化后结果 $\tilde{\mathbf{X}}$

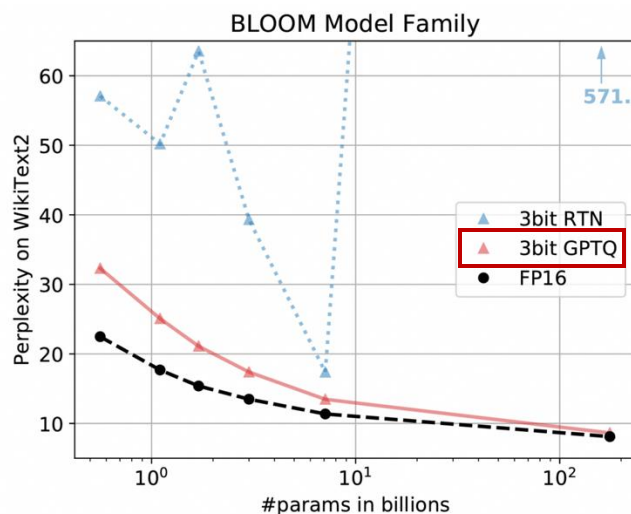
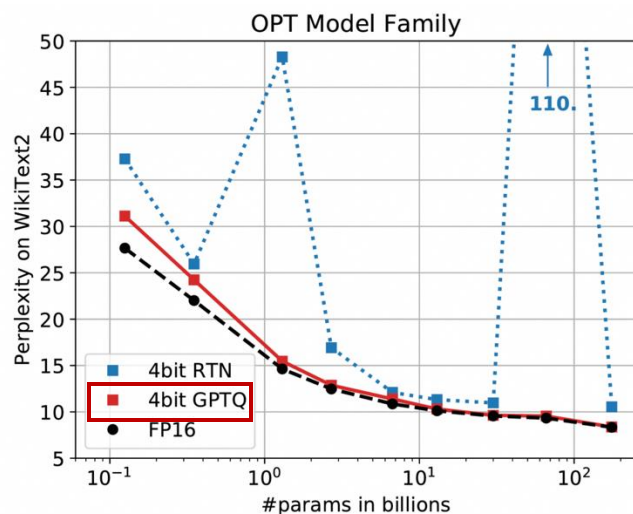
训练后量化方法

➤ 权重量化 (W 是原始权重)

➤ 最小化重构损失 (W_q 是量化后权重)

$$\arg \min_{W_q} \|XW - XW_q\|_2^2$$

➤ GPTQ: 将权重矩阵按照列维度分组, 逐组量化

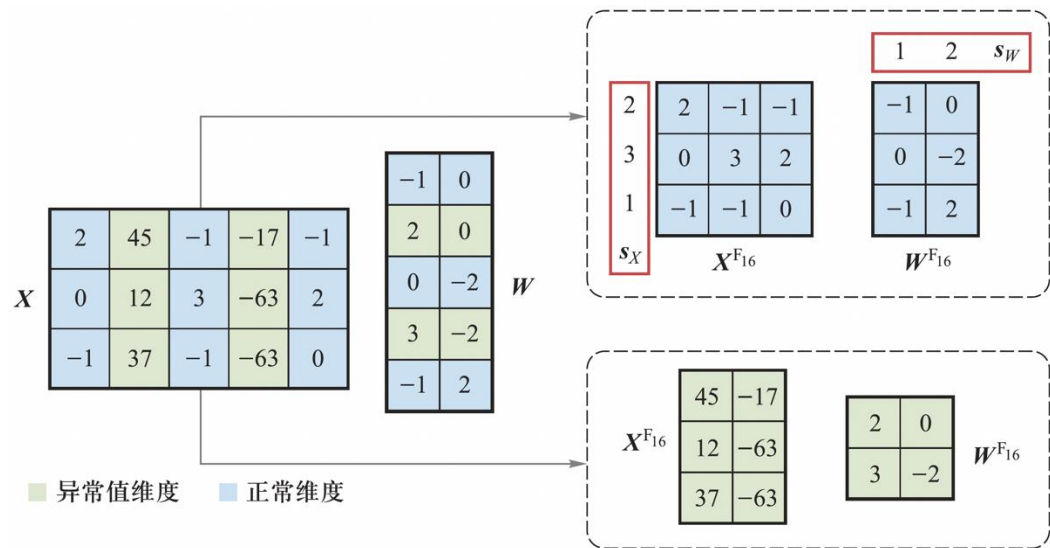
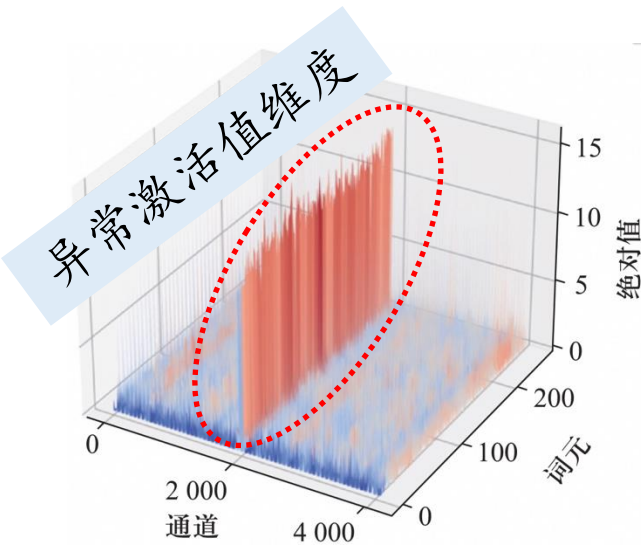
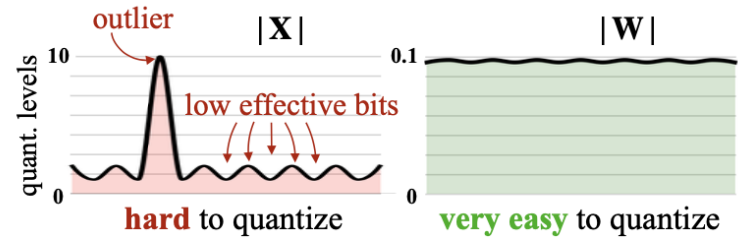


3/4 比特量化与16比特对比

- 模型越大, 表现越接近

➤ 权重和激活值量化

- 模型达到一定规模，某些维度会出现异常激活值
- 混合精度分解：将异常值（16比特）和正常值（8比特）分开计算



正常值使用8
比特量化

异常值使用16
比特计算

➤ 模型蒸馏

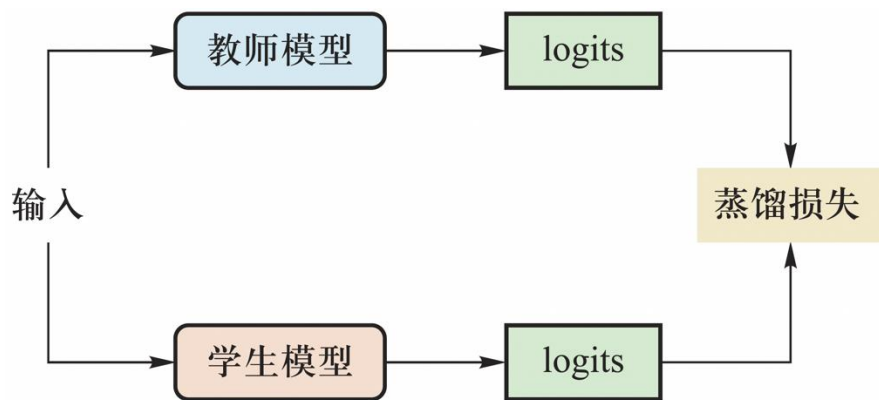
➤ 将复杂模型的知识迁移到简单模型上

➤ 基于反馈的知识蒸馏

➤ 使用教师模型的输出概率分布作为学生模型的“软标签”

$$\mathcal{L}(l_t, l_s) = \mathcal{L}_R(P_t(\cdot), P_s(\cdot))$$

\mathcal{L}_R 是损失函数
常用KL散度



让学生模型的输出
与教师模型接近

➤ 模型蒸馏

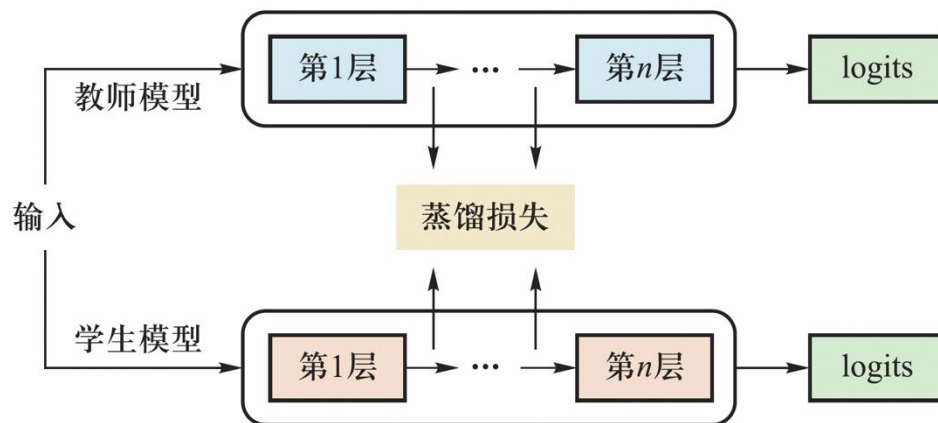
➤ 将复杂模型的知识迁移到简单模型上

➤ 基于特征的知识蒸馏

➤ 使用教师模型中间层的输出特征作为监督信息训练学生模型

$$\mathcal{L}(f_t(\mathbf{x}), f_s(\mathbf{x})) = \mathcal{L}_F(\Phi(f_t(\mathbf{x})), \Phi(f_s(\mathbf{x})))$$

f 是中间层输出特征
 $\Phi(\cdot)$ 用于转换输出维度



相较于输出层，中间层
能提供更丰富的信息



谢谢