模型压缩

《大语言模型》编写团队：唐天一

量化基础知识

AI Box

- 量化：将映射浮点数到整数的过程

$$\boldsymbol{X}_q = R(\boldsymbol{X}/S) - Z$$

R：取整函数

S：放缩因子

Z：零点因子

- 反量化：从量化值中恢复原始值

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

- 量化误差：原始值 $\boldsymbol{X}$ 和恢复值 $\tilde{\boldsymbol{X}}$ 之间的数值差异

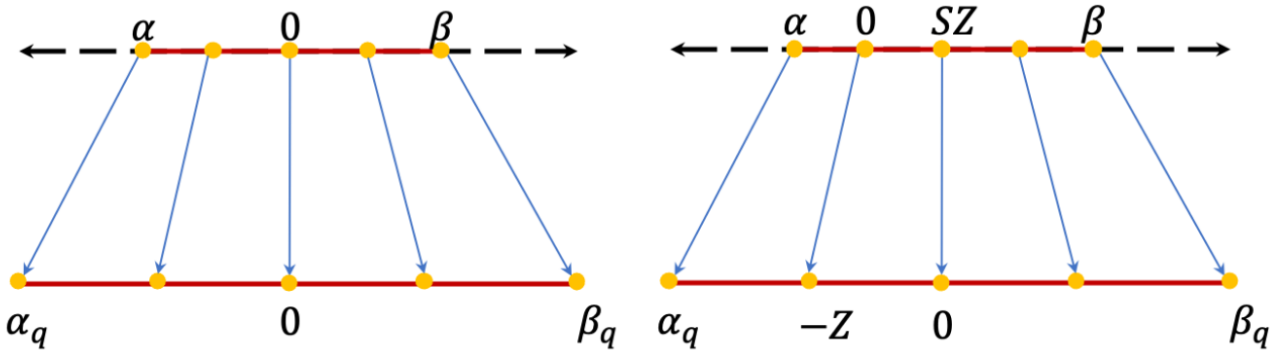
$$\Delta = \left\| \boldsymbol{X} - \tilde{\boldsymbol{X}} \right\|_2^2$$

量化的目标是最小化量化误差

量化基础知识

AI Box

- 对称量化和非对称量化
  - 根据零点因子 Z 是否为零



对称量化（$Z=0$）　　非对称量化（$Z\neq 0$）

- 量化粒度的选择

-1	0
0	-2
-1	2

$\boldsymbol{W}_{F16}$

2

$\boldsymbol{s}_W$

张量量化：一个矩阵定义一组量化参数

1	2	$\boldsymbol{s}_W$
-1	0	
0	-2	
-1	2	

$\boldsymbol{W}_{F16}$

通道量化：一个矩阵对列维度设置特定的量化参数

量化基础知识

AI Box

- 非对称量化方法计算示例

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

确定输入范围 $\in [1.2, 13.6]$

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

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

量化到整数范围 $[-128, 127]$ 两个边界值映射

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

计算量化参数 S 和 Z

$$\boldsymbol{X_q} = [[-127,-103,-78],[78,103,127]]$$

计算量化后结果 $\boldsymbol{X_q}$

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

计算反量化后结果 $\tilde{\boldsymbol{X}}$

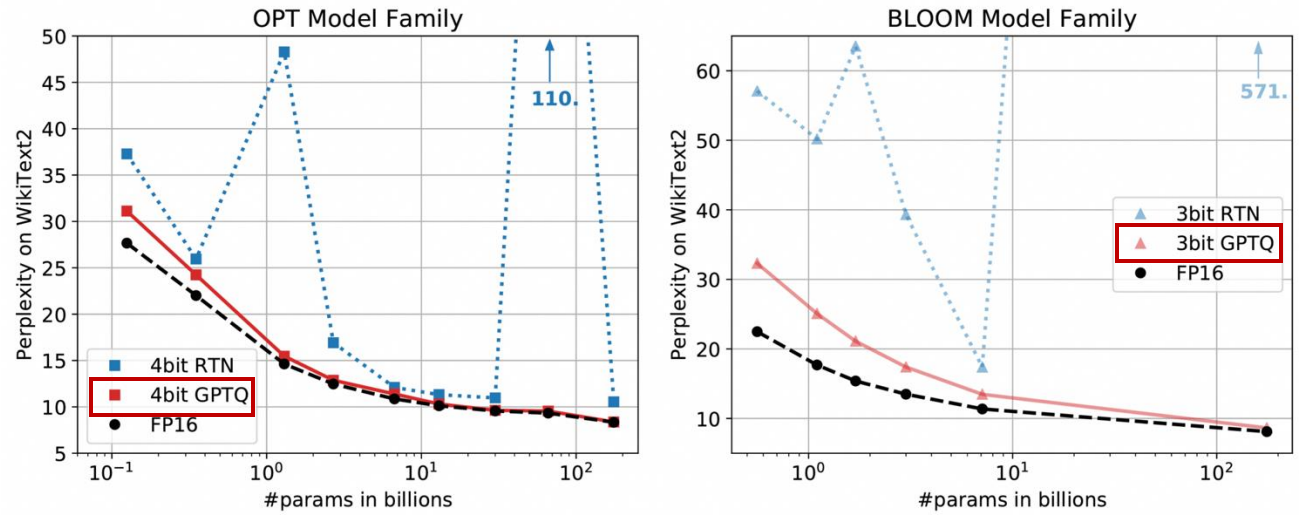
训练后量化方法

AI Box

- 权重量化（$\boldsymbol{W}$ 是原始权重）
  - 最小化重构损失（$\boldsymbol{W}_q$ 是量化后权重）

$$\arg\min_{\boldsymbol{W}_q} \| \boldsymbol{X}\boldsymbol{W} - \boldsymbol{X}\boldsymbol{W}_q \|_2^2$$

  - GPTQ：将权重矩阵按照列维度分组，逐组量化



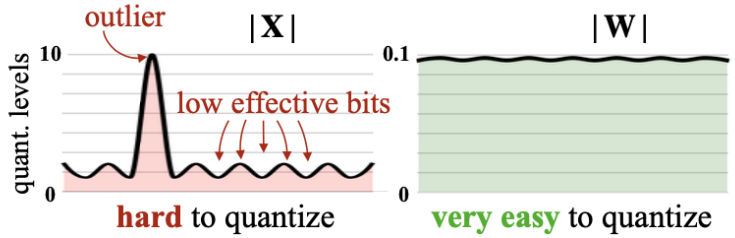
3/4 比特量化与16比特对比

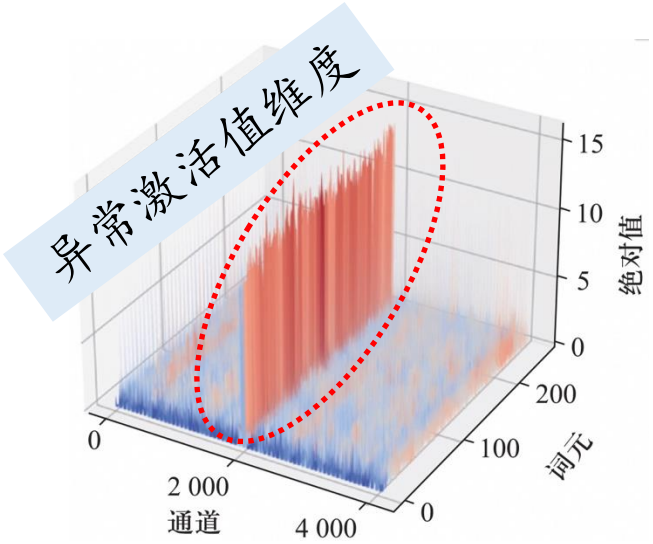
- 模型越大，表现越接近

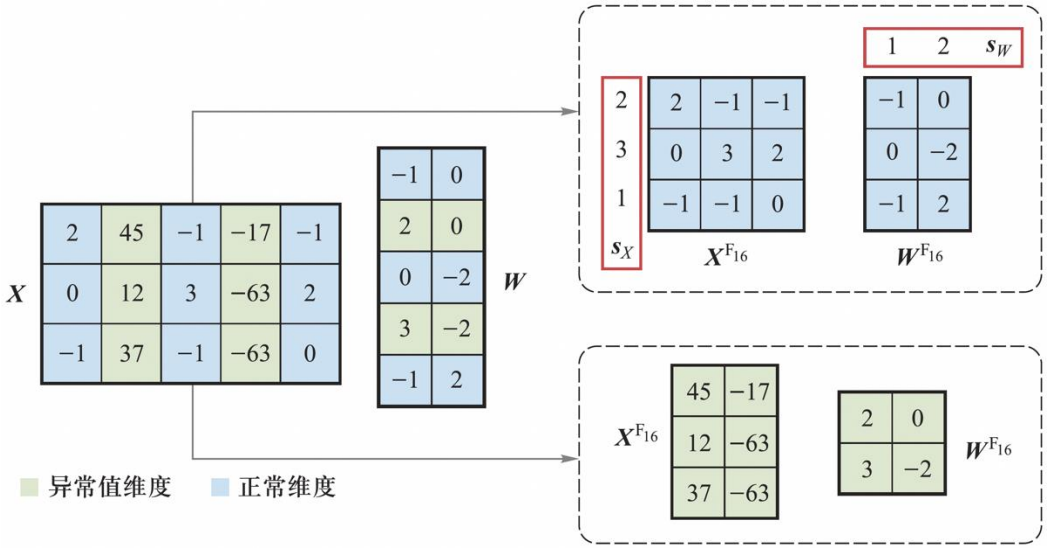
GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers, **ICLR 2023**

训练后量化方法

- 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

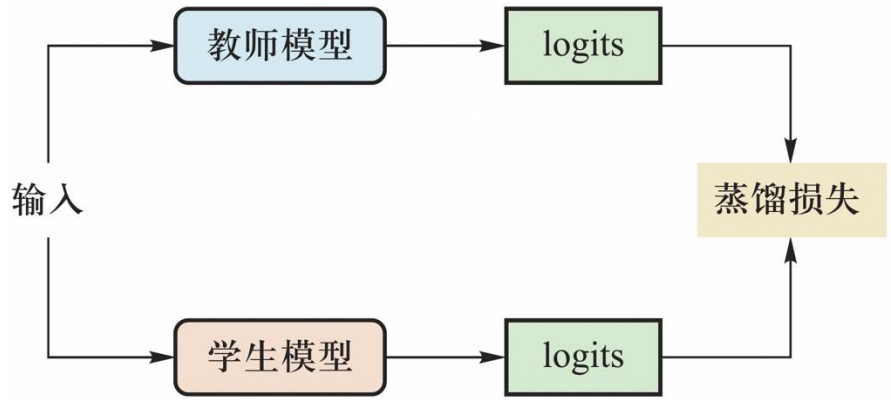
模型蒸馏

- 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”

$$\mathcal{L}(\boldsymbol{l}_t, \boldsymbol{l}_s) = \mathcal{L}_R(P_t(\cdot), P_s(\cdot))$$

$\mathcal{L}_R$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

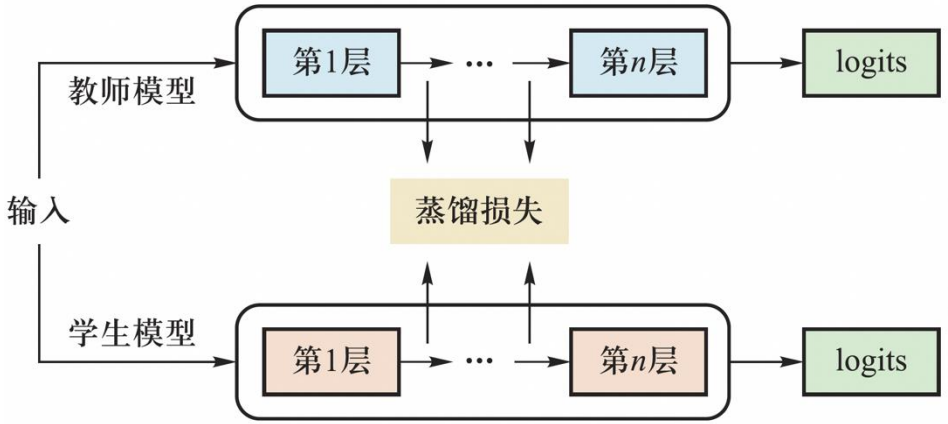
模型蒸馏

- 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

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

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