# PTQ：训练后量化

# 一，动态离线量化

动态离线量化，将模型中特定OP的权重从FP32类型量化成INT8/16类型。

## 两种预测方式：

* 第一种是反量化预测方式，即是首先将INT8/16类型的权重反量化成FP32类

型，然后再使用FP32浮运算运算进行预测

* 第二种量化预测方式，即是预测中动态计算量化OP输入的量化

信息，基于量化的输入和权重进行INT8整形运算

## 使用条件：

* 有训练好的预测模型

* 权重量化成INT16类型，模型精度不受影响，模型大小为原始的1/2。

* 权重量化成INT8类型，模型精度会受到影响，模型大小为原始的1/4。

## 优点：

* 权重量化成INT16类型，模型精度不受影响，模型大小为原始的1/2

* 权重量化成INT8类型，模型精度会受到影响，模型大小为原始的1/4

## 缺点：

* 目前PaddleLite只支持反量化预测方式，主要可以减小模型大小，对特定加载权重费时的模型可以起到一定加速效果

## 原理：

其工作比较简单，在端测转换工具的时候，对网络模型的权重进行统计其每一层卷积的layer或者channel的最大值和最小值，然后通过量化公式对数据进行byte转换。这样得到的权重参数比以前小1/4。推理的时候，在内存初始化的时候对网络模型中的权重进行反量化操作变成float进行正常的推理。

## PaddleSlim接口：

**paddleslim.quant.quant\_post\_dynamic(*model\_dir, save\_model\_dir, model\_filename=None, params\_filename=None, save\_model\_filename=None, save\_params\_filename=None, quantizable\_op\_type=["conv2d", "mul"], weight\_bits=8, generate\_test\_model=False*)**

***其内部使用channel\_wise\_abs\_max方法进行量化，其zero\_point/offset=0***

# 二，静态离线量化

静态离线量化是基于采样数据，采用KL散度等方法计算量化比例因子的方法。相比量化训练，静态离线量化不需要重新训练，可以快速得到量化模型。

静态离线量化，使用少量校准数据计算量化因子，可以快速得到量化模型。使用该量化模型进行预测，可以减少计算量、降低计算内存、减小模型大小。

## 使用条件:

* 有训练好的预测模型

* 有少量校准数据，比如几十到几百张图片

## 优点：

* 减小计算量、降低计算内存、减小模型大小

* 不需要大量训练数据

* 快速产出量化模型，简单易用

## 缺点：

* 对少部分的模型，尤其是计算量小、精简的模型，量化后精度可能会受到影响

## 原理：

### 1.权重量化

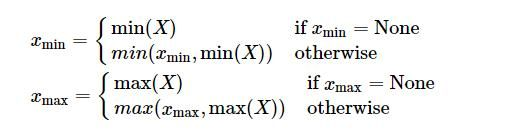
权值在进行推理加速时均已确定，因此不需要对权值进行校准。如果使用对称量化方案，max使用权值的绝对值的最大值，对于非对称量化算法，max和min使用权值的最大值和最小值。

### 2.激活量化：

激活的实际动态范围则必须经过采样获取(一般把这个过程称为数据校准(calibration) )。目前各个深度学习框架中，使用最多的有最大最小值(MinMax)， 滑动平均最大最小值(MovingAverageMinMax) 和 KL 距离(Kullback–Leibler divergence) ，abs\_max，avg，mse，hist。如果量化过程中的每一个 FP32 数值都在这个实际动态范围内，我们一般称这种为不饱和状态；反之如果出现某些 FP32 数值不在这个实际动态范围之内我们称之为饱和状态。

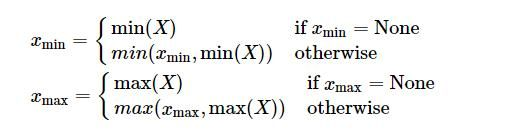
#### 滑动平均最大最小值：

会采用一个超参数 c 逐步更新动态范围。这种方法获得的动态范围一般要小于实际的动态范围。对于 weights 而言，由于不存在采样的迭代，因此 MovingAverageMinMax 与 MinMax 的效果是一样的。



#### 最大最小值：

这是最简单也是使用比较多的一种采样方法。它的基本思想是直接从 FP32 张量中选取最大值和最小值来确定实际的动态范围。



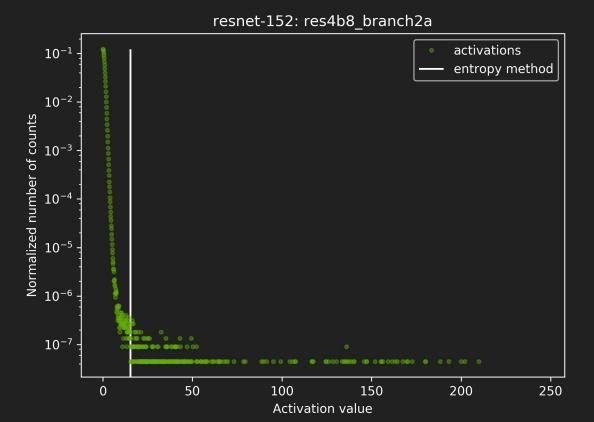
#### KL 距离采样方法(Kullback–Leibler divergence)：

量化是对原始 FP32数据的一种重新编码。一般认为量化之后的数据分布与原始分布越相似，量化对原始数据信息的损失也就越小。KL 距离一般被用来度量两个分布之间的相似性。



其中P，Q表示两个不同的分布。

动态范围的选取直接决定了量化数据的分布情况，处于动态范围之外的数据将被映射成量化数据的边界点。如下图所示，横坐标表示activation 的取值，纵坐标表示每个取值的归一化统计个数。从图可以看出绝大部分数值都分布在白色直线的左端。通过 KL 距离采样方法就会将动态范围限制在白线左侧的部分，白线右边的值将都会被映射成量化数据的最大值。



## PaddleSlim接口:

**paddleslim.quant.quant\_post\_static(*executor,model\_dir, quantize\_model\_path, batch\_generator=None, sample\_generator=None, model\_filename=None, params\_filename=None, save\_model\_filename='\_\_model\_\_', save\_params\_filename='\_\_params\_\_', batch\_size=16, batch\_nums=None, scope=None, algo='KL', quantizable\_op\_type=["conv2d","depthwise\_conv2d","mul"], is\_full\_quantize=False, weight\_bits=8, activation\_bits=8, activation\_quantize\_type='range\_abs\_max', weight\_quantize\_type='channel\_wise\_abs\_max', optimize\_model=False*)**

***内部实现查看“post\_training\_quantization.py”***