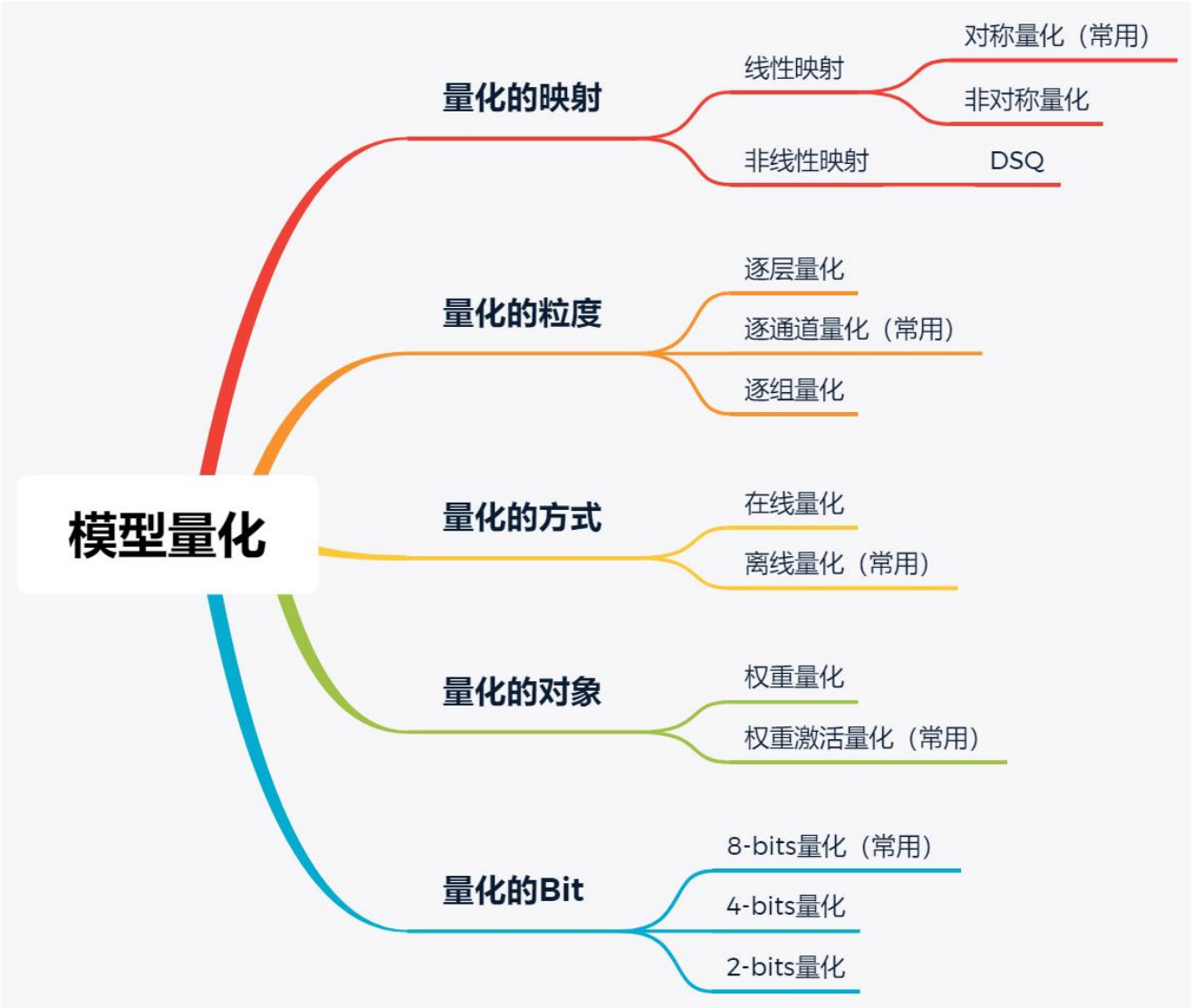
模型量化算法


模型量化
量化的映射
线性映射
对称量化（常用）
非对称量化
非线性映射
DSQ
量化的粒度
逐层量化
逐通道量化（常用）
逐组量化
量化的方式
在线量化
离线量化（常用）
量化的对象
权重量化
权重激活量化（常用）
量化的Bit
8-bits量化（常用）
4-bits量化
2-bits量化

1、仿射映射量化

至于模型的量化映射方式，如上图所示可以分为线性映射和非线性映射，对于线性映射又可以分为对称量化和非对称量化。

由于实际项目中基本不会使用非线性量化，本次课程不对非线性量化进行介绍。

2、线性映射

对称量化：

对称量化即使用一个映射公式将输入数据映射到 [-127,127] 的范围内映射公式需要保证原始的输入数据中的零点通过映射公式后仍然对应 [-127, 127] 区间的零点。

量化

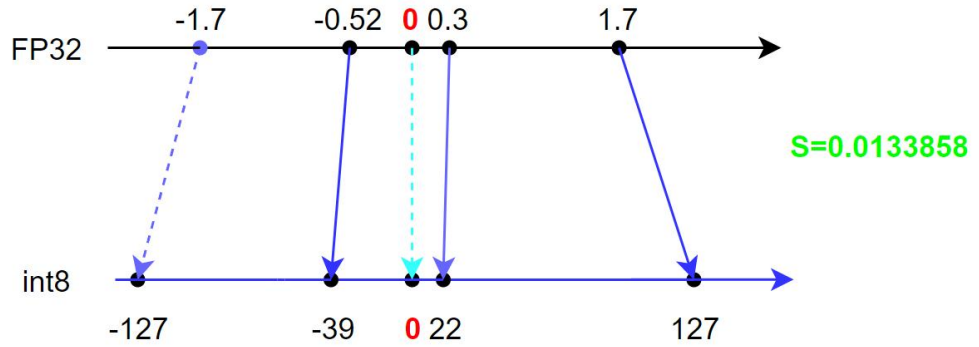
$$Q = Round\left(\frac{R}{Scale}\right)$$

$$Scale = \frac{|R_{max}|}{|Q_{max}|}$$

反量化

$$R = Q \times Scale + Z$$

$$Scale = \frac{|R_{max}|}{|Q_{max}|}$$



非对称量化：

即使用一个映射公式将输入数据映射到[-128,127] 的范围内。但是原始的输入数据中的零点通过映射公式后对应的位置并不是原点。

量化

$$Q = Round\left(\frac{R}{Scale} + Z\right)$$

$$Scale = \frac{R_{max} - R_{min}}{Q_{max} - Q_{min}}$$

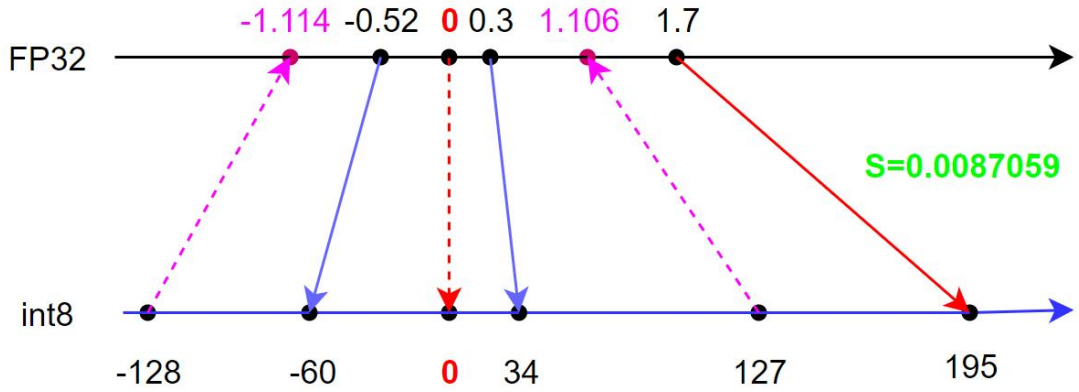
$$Z = Q_{max} - Round\left(\frac{R_{max}}{Scale}\right)$$

反量化

$$R = Q \times Scale + Z$$

$$Scale = \frac{R_{max} - R_{min}}{Q_{max} - Q_{min}}$$

$$Z = Q_{max} - Round\left(\frac{R_{max}}{Scale}\right)$$



3、逐层量化、逐组量化和逐通道量化

根据量化的粒度可以分为**逐层量化（per-Tensor）**、**逐组量化（per-Group）**和**逐通道量化（per-Channel）**。

- **逐层量化**以一个层为单位，整个 layer 的权重共用一组缩放因子 S 和偏移量 Z;
- **逐组量化**以组为单位，每个 Group 使用一组 S 和 Z;
- **逐通道量化**则以通道为单位，每个 channel 单独使用一组 S 和 Z。

当 Group =1 时，逐组量化与逐层量化等价；当 Group =num_filters （即深度可分离卷积）时，逐组量化逐通道量化等价。

4、在线量化和离线量化

根据激活值的量化方式，可以分为在线（online）量化和离线（offline）量化。

- **在线量化**：指激活值的Scale和Z-point在实际推断过程中根据实际的激活值动态计算；

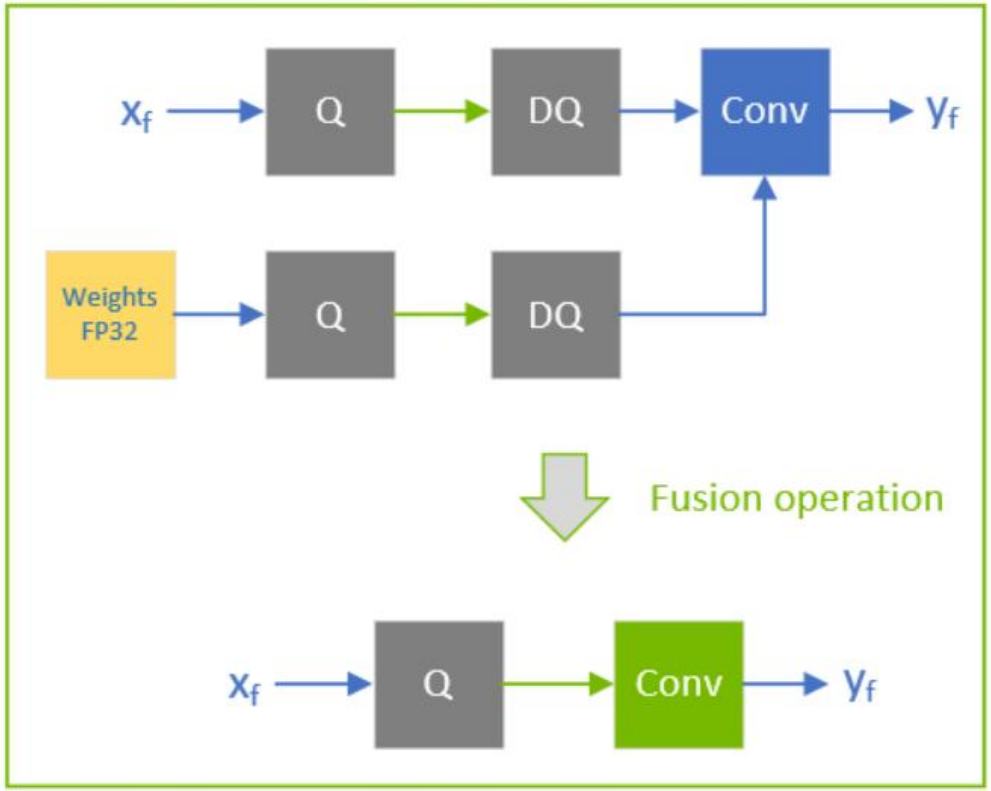
- **离线量化**：即指提前确定好激活值的Scale和 Z-point 。

由于不需要动态计算量化参数，通常离线量化的推断速度更快些。

5、权重量化和权重激活量化

权重量化：只将权重的精度从浮点型减低为8bit整型。由于只有权重进行量化，所以无需验证数据集就可以实现。如果只是想为了方便传输和存储而减小模型大小，而不考虑在预测时浮点型计算的性能开销的话，这种量化方法是很有用的。

权重激活量化：可以通过计算所有将要被量化的数据的量化参数，来将一个浮点型模型量化为一个8bit精度的整型模型。由于激活输出需要量化，这时我们就得需要标定数据了，并且需要计算激活输出的动态范围，一般使用100个小批量数据就足够估算出激活输出的动态范围了。



6、量化的一般步骤

对于模型量化任务而言，具体的执行步骤如下所示：

步骤1：在输入数据（通常是权重或者激活值）中计算出相应的 min_value 和 max_value；

步骤2：选择合适的量化类型，**对称量化**（int8）还是非对称量化（uint8）；

步骤3：根据量化类型、min_value 和 max_value 来计算Z（Zero point） 和 S（Scale）；

步骤4：根据标定数据对模型执行量化校准操作，即将预训练模型由 FP32 量化为 INT8模型；

步骤5：验证量化后的模型性能，如果效果不好，尝试着使用不同的方式计算 S 和 Z，重新执行上面的操作。

步骤6：如果步骤5.未能满足性能要求，则尝试使用敏感层分析和量化感知训练。