模型优化与部署

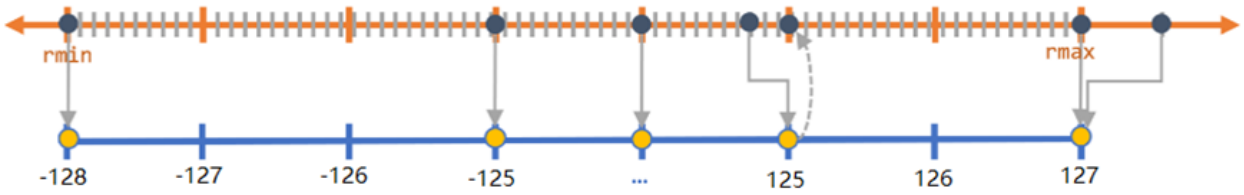
量化

概念

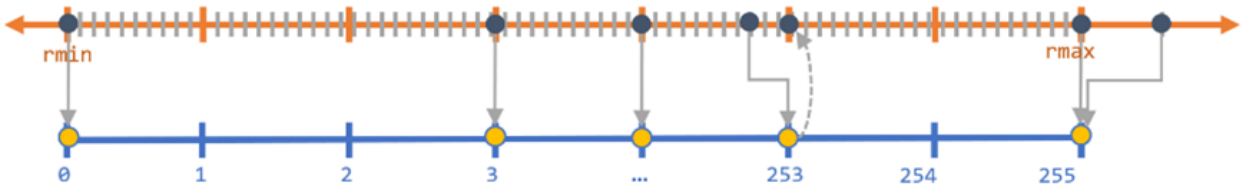
简单理解为用低精度数来近似表示高精度浮点数（网络权值、激活值等）

对称量化

以int8为例，将最大绝对值映射到8位的最大值，最大绝对值的负数映射到8比特的最小值



非对称量化



通过收缩因子和零点，将FP32张量的min/max分别映射到8-bit数据的min/max

线性量化和非线性量化

线性量化被称为均匀量化，非对称量化和对称量化都是基于线性量化的，较为常用；非线性量化应用较少，主要以LOG量化为代表

不同的量化方式对数据分布具有选择性。对于均匀量化，假设数据在整个表达空间内均匀分布，在均匀分布下线性量化是一种较好的量化方式。 LOG 量化则可以保证数值空间内相对误差的最优化，这是目前大部分非线形量化的目标，通过对数据分布的分析，可以提升高密度数据区域的表达能力。

量化策略

如果知道了阈值，那么其对应的线性映射参数也就知道了，整个量化过程也就明确了。那么该如何确定阈值呢？一般来说，对于权重的量化，由于权重的数据分布是静态的，一般直接找出 MIN 和 MAX 线性映射即可；而对于**推理激活值**来说，其**数据分布是动态的**，为了得到激活值的数据分布，往往需要一个所谓**校准集**的东西来进行抽样分布，有了抽样分布后再通过一些量化算法进行量化阈值的选取

MinMax 量化：

MinMax 其实就是简单的把浮点数直接映射到 int8 的数据范围

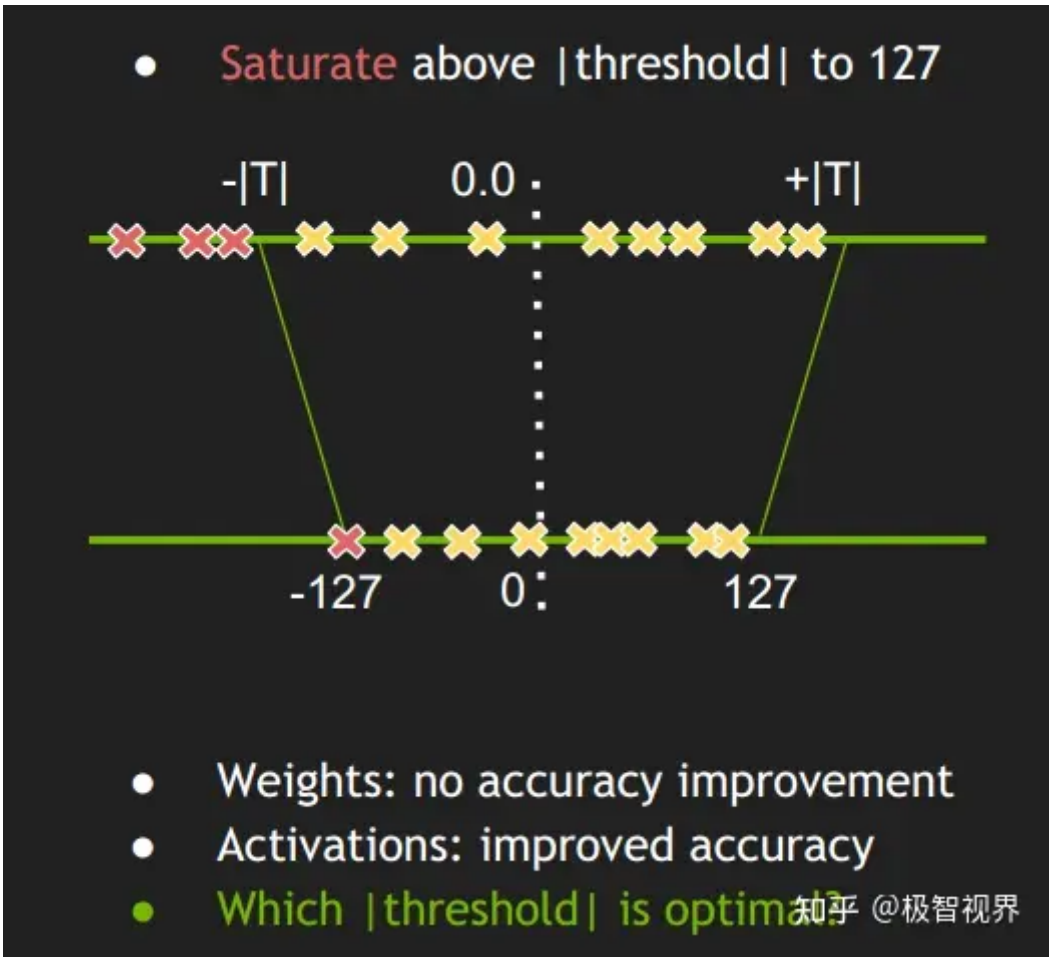
$$Q = \frac{R}{S} + Z, \quad S = \frac{Rmax - Rmin}{Qmax - Qmin}, \quad Z = Qmax - \frac{Rmax}{S}$$

其中，R：真实浮点值(fp32)；Q：量化后的定点值(int8，Q属于[-127, 127])；Z：表示0浮点值对应的量化定点值；S：定点量化后可表示的最小刻度

把 MinMax 量化应用于网络权重这样静态分布的数据的时候，对于网络推理最后的精度损失影响不大，且量化操作开销更小，量化过程效率更高，这是 Nvidia 经过大量实验得出的结论。

KLD 量化

KLD 量化是用 KL 散度来衡量两个分布之间的相似性



- 这种方法不是直接将 [min, max] 映射到 [-127,127]，而是去寻找一个阈值|T| < max(|max|, |min|)，将其 [-T, T] 映射到 [-127, 127]。认为只要阈值选取得当，就能将阈值外的值舍弃掉，也不会对精度损失造成大的影响；
- 超出阈值 ±|T| 外的直接映射为阈值。比如上图中的三个红色点，直接映射为-127

滑动平均最大最小值

与 `MinMax` 算法直接替换不同，MovingAverageMinMax 会采用一个超参数 `c`

$$x_{min} = \begin{cases} min(X) & if x_{min} = None \\ (1-c)x_{min} + c\,min(X) & otherwise \end{cases}$$

$$x_{max} = \begin{cases} max(X) & if x_{max} = None \\ (1-c)x_{max} + c\,max(X) & otherwise \end{cases}$$

ADMM

EQ

EQ 量化方法的主要思想是：误差累计、整网决策变成单网决策、以余弦相似度为优化目标、交替优化权重缩放系数和激活值缩放系数。

假设量化公式为：

$$Q(X,S) = clip(round(X * S + Z))$$

量化精度为intN,则：

$$Q(X,S) \in \{ x \in Z | -2^{n-1} <= x <= -2^{n-1} - 1\}$$

Oil 表示未量化推理时第 l 层网络层的第 i 个样本的输出值，O^il 表示量化推理时第 l 层网络层的第 i 个样本的输出值。

EQ算法流程：

输入：模型权重集合$\{W_l\}_{l=1}^{L}$，根据校准集在原始模型中推理保存下来的每层的输入激活值和输出的集合$\{A_l, O_l\}_{l=1}^{L}$

1. min-max 初始化每层的缩放系数$\{S_l\}_{l=1}^{L}$
2. while 未达到条件：
3. 固定激活值缩放系数S^a
4. for l in range(1, L+1):
5. $$S_l^w = \arg\max_{S_l^w \in [\alpha S_l^w, \beta S_l^w]} \sum_{i=1}^{N} \cos(O_l^i, \hat{O}_l^i)$$
6. 固定权重缩放系数S^w
7. for l in range(1, L+1):
8. $$S_l^a = \arg\max_{S_l^a \in [\alpha S_l^a, \beta S_l^a]} \sum_{i=1}^{N} \cos(O_l^i, \hat{O}_l^i)$$
9. 根据更新得到的缩放系数$\{S_l\}_{l=1}^{L}$对整网推理，每层的输入使用量化的累计误差得到$\hat{A}_l$
10. 返回满足条件的缩放系数$\{S_l\}_{l=1}^{L}$

参考：

https://zhuanlan.zhihu.com/p/414647262