**基本量化公式**

r 表示浮点实数，q 表示量化后的定点整数。浮点和整型之间的换算公式为：

其中，S 是 scale，表示实数和整数之间的比例关系，Z是 zero point，表示实数中的 0 经过量化后对应的整数。

## 矩阵的量化原理

由于卷积网络中的卷积层和全连接层本质上都是一堆矩阵乘法，因此我们先看如何将浮点运算上的矩阵转换为定点运算。

假设 r1、r2 是浮点实数上的两个N×N的矩阵，r3 是r1、r2 相乘后的矩阵：

假设S1、Z1 是R1矩阵对应的 scale 和 zero point，S2 、 Z2 、 S3 、 Z3 同理，那么可以推出：

整理一下可以得到：

仔细观察可以发现，除了，其他都是定点整数运算。

那如何把也变成定点运算呢？这里要用到一个trick。

假设 ，由于M通常都是 (0, 1) 之间的实数 (这是通过大量实验统计出来的)，因此可以表示成，其中是一个定点实数。

定点数并不一定是整数，所谓定点，指的是小数点的位置是固定的，即小数位数是固定的。因此，如果存在，那我们就可以通过的 bit 位移操作实现 ，这样整个过程就都在定点上计算了。

## 卷积的量化原理

卷积运算和全连接层的本质都是矩阵运算，因此我们可以把卷积运算表示成矩阵运算。

事先统计并计算:

1.input的(scale)和(zero point)

2.weigh的(scale)和(zero point)

3.bias的(scale)和(zero point)

4.output的(scale)和(zero point)

可以得到：

在咱们的实际项目中，zero point都为0，因此可以简化：

简化后为：

可得到：

其中，，在实际项目中为了简化计算，因此进行如下调整：

令=1，这样就可以直接定点相乘后直接相加：

因此存在如下约束：

=

进一步检验实际case的结果：某卷积Conv7中的量化数据如下:

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

核对M的大小是否可以符合前文约束：

核对的大小是否可以符合前文约束：

## 卷积计算模拟

在计算卷积操作时，使用以下公式来计算输出的尺寸：

Output size

Inputsize 是输入的尺寸；Fitersize是卷积核的尺寸；

Padding是填充大小，Stride是步长。

卷积计算中，输入input、卷积kernel、偏置bias的维度通常用"nwhc"表示，其中：

* n: 批处理大小batch size
* w: 输入宽度width
* h: 输入高度height
* c: 输入通道数channel

假如维度如下：

* 输入input：1 \* 7 \* 10 \* 128
* 卷积核kernel：128 \* 3 \* 3 \* 128
* 偏置bias：1 \* 1 \* 1 \* 128

在卷积操作中，输出的维度计算如下：

* output\_height\_height\_kemel height+2xpadding height output channel=number of kernels 根据你提供的信息： output\_ width output\_height\_ output\_channel\_128
* 因此，输出 (output) 的维度为： ம்மு 5

Output size

2. 计算卷积操作：

对于每个输出通道和每个位置:

然后加上偏置项：

Output

在本案例中的没有填充。

## 卷积网络的误差分析