FLOPs

> **这里先注意一下FLOPs的写法，不要弄混了：**
> **FLOPS(全大写)**：是floating point operations per second的缩写，意指每秒浮点运算次数，理解为计算速度，是一个衡量硬件性能的指标。
> **FLOPs(s小写)**：，是floating point operations的缩写（s表复数），意指浮点运算数，理解为计算量，可以用来衡量算法/模型的复杂度，也就是我们这里要讨论的。

标准卷积层的FLOPs

考虑bias：$(2 * C_{int} * k^2) * C_{out} * H * W$
不考虑bias：$(2 * C_{int} * k^2 - 1) * C_{out} * H * W$

参数定义（下同）：C_{int}为输入通道数，k为卷积核边长，C_{out}为输出通道数，H*W为输出特征图的长宽。

其实卷积层在实现的时候可以选择加bias或者不加，在很多的框架当中是一个可以选择的参数，为了严谨，这里特地提一下。

怎么理解上面的公式呢？以不考虑bias为例。我们先计算输出的feature map中的一个pixel的计算量，然后再乘以feature map的规模大小即可，所以我们主要分析下上面公式中的括号部分：

$$(2 * C_{int} * k^2 - 1) = C_{int} * k^2 + C_{int} * k^2 - 1 \tag{1}$$

可以看到我们把它分成了两部分，**第一项是乘法运算数，第二项是加法运算数，因为n个数相加，要加n-1次，所以不考虑bias，会有一个-1，如果考虑bias，刚好中和掉。**

深度可分离卷积的FLOPs

深度可分离卷积分成两部分，一部分是分通道卷积，另一部分是1*1卷积。（如果不知道深度可分离卷积的朋友可以先看下这个博客，这是一种可大大减少计算量的卷积方法）

这里的讨论以考虑bias为准：
第一部分：$(2 * k^2) * H * W * C_{int}$
第二部分：$2 * C_{int} * H * W * C_{out}$

最终的结果就是两部分相加。

池化层的FLOPS

这里又分为全局池化和一般池化两种情况：

全局池化

针对输入所有值进行一次池化操作，不论是max、sum还是avg，都可以简单地看做是只需要对每个值算一次。

所以结果为：$H_{int} * W_{int} * C_{int}$

一般池化

答案是：$k^2 * H_{out} * W_{out} * C_{out}$

注意池化层的：$C_{out} = C_{int}$

全连接层的FLOPs

考虑bias：$(2*I)*O$
不考虑bias：$(2*I-1)*O$

分析同理，括号内是一个输出神经元的计算量，拓展到O个输出神经元。（如果该全连接层的输入是卷积层的输出，需要先将输出展开成一列向量）

激活层的FLOPs

ReLU

ReLU一般都是跟在卷积层的后面，这里假设卷积层的输出为$H*W*C$，因为ReLU函数的计算只涉及到一个判断，因此计算量就是$H*W*C$

sigmoid

根据sigmoid的公式可以知道，每个输入都需要经历4次运算，因此计算量是$H*W*C*4$（参数含义同ReLU）

参数量

卷积层的参数量

卷积层的参数量与输入特征图大小无关

考虑bias：$(k^2*C_{int}+1)*C_{out}$
不考虑bias：$(k^2*C_{int})*C_{out}$

深度可分离卷积的参数量

不考虑bias：
第一部分：k^2*C_{int}
第二部分：$(1*1*C_{int})*C_{out}$
最终结果为两者相加。

池化层的参数量

池化层没有需要学习的参数，所以参数量为0。

全连接层的参数量

考虑bias：$I*O+1$

如果有写的不对的地方请各位大佬在评论区指出来，一起学习！

参考资料

CNN 模型所需的计算力（flops）和参数（parameters）数量是怎么计算的？