# 深度学习模型的性能分析

## 参考文献

https://zhuanlan.zhihu.com/p/34204282

## 台的两个指标：算力  与带宽

1. **算力：**也称为计算平台的性能上限，指的是一个计算平台倾尽全力每秒钟所能完成的浮点运算数。单位是FLOP/s。

 Maximum Flops Per Second

1. **带宽：**也即计算平台的带宽上限，指的是一个计算平台倾尽全力每秒所能完成的内存交换量。单位是Byte/s。

**：**Maximum Memory Access Per Second

1. **计算强度上限 :** 两个指标相除即可得到计算平台的计算强度上限。它描述的是在这个计算平台上，单位内存交换最多用来进行多少次计算。单位是Flop/Byte: **；**

## 两个指标：计算量 与 访存量

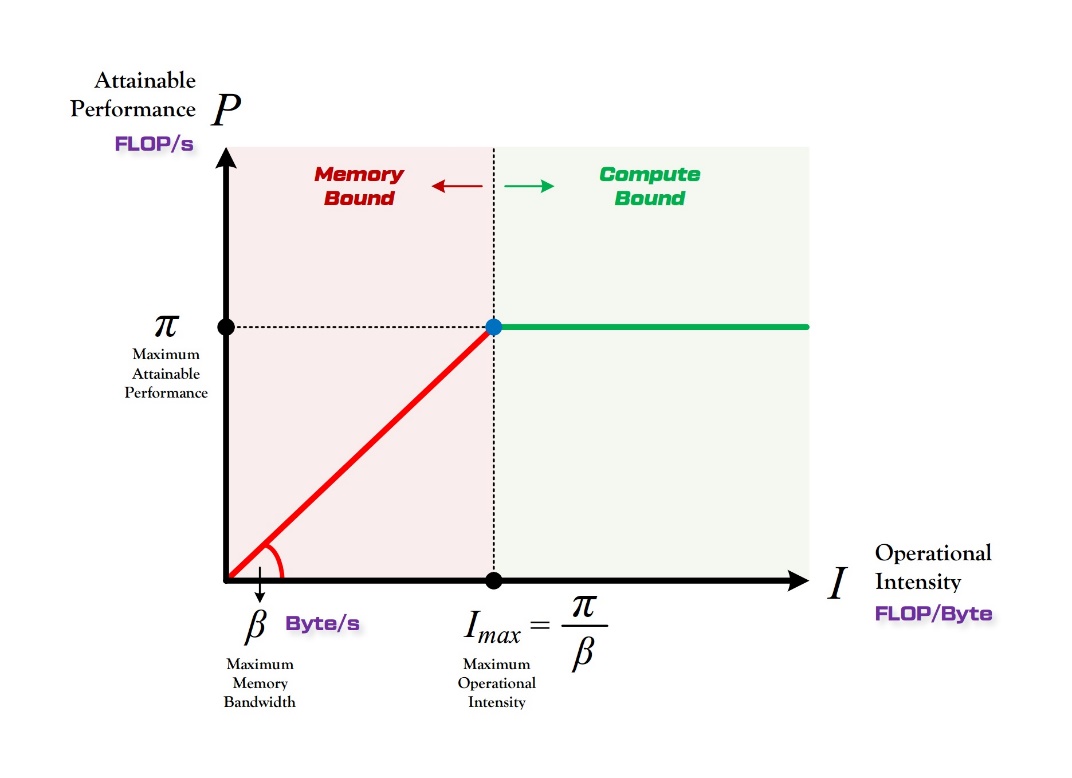
1. **计算量：**指的是输入单个样本（对于CNN而言就是一张图像），模型进行一次完整的前向传播所发生的浮点运算个数，也即模型的时间复杂度。单位是FLOPS;
2. **访存量：**指的是输入单个样本，模型完成一次前向传播过程中所发生的内存交换总量，也即模型的空间复杂度。在理想情况下（即不考虑片上缓存），模型的访存量就是模型各层权重参数的内存占用与每层所输出的特征图的内存占用之和。单位是Byte。由于数据类型通常为float32 ，因此需要乘以四。

Conv Layer Space Complexity: 

1. **模型的计算强度:**由计算量除以访存量就可以得到模型的计算强度，它表示此模型在计算过程中，每Byte内存交换到底用于进行多少次浮点运算。单位FLOP/Byte。可以看到，模计算强度越大，其内存使用效率越高。
2. **模型的理论性能P:**我们最关心的指标，即模型在计算平台所能达到的每秒浮点运算次数。单位是FLOP/S。

## Roof-line Model

* 1. 所谓“Roof-line”，指的就是由计算平台的算力和带宽上限这两个参数所决定的“屋顶”形态，如下图所示。
* **算力**决定“屋顶”的高度（绿色线段）
* **带宽**决定“房檐”的斜率（红色线段）



## Roof-line 划分出的两个瓶颈区域：

**计算瓶颈区域 Compute-Bound:** 不管模型的计算强度有多大，它的理论性能 P最大只能等于计算平台的算力 。当模型的计算强度  大于计算平台的计算强度上限  时，模型在当前计算平台处于 Compute-Bound状态，即模型的理论性能 P 受到计算平台算力  的限制，无法与计算强度  成正比。但这其实并不是一件坏事，因为从充分利用计算平台算力的角度上看，此时模型已经  的利用了计算平台的全部算力。可见，计算平台的算力  越高，模型进入计算瓶颈区域后的理论性能 P 也就越大。

**带宽瓶颈区域 Memory-Bound:** 当模型的计算强度  小于计算平台的计算强度上限  时，由于此时模型位于“房檐”区间，因此模型理论性能 P 的大小完全由计算平台的带宽上限  （房檐的斜率）以及模型自身的计算强度  所决定，因此这时候就称模型处于 Memory-Bound 状态。可见，在模型处于带宽瓶颈区间的前提下，计算平台的带宽  越大（房檐越陡），或者模型的计算强度  越大，模型的理论性能 P 可呈线性增长。