

该图是AlexNet网络中不同层的GPU和CPU的时间消耗，我们可以清晰的看到，不管是在GPU还是在CPU运行，最重要的“耗时杀手”就是conv，卷积层。也就是说，想要提高网络的运行速度，就得到提高卷积层的计算效率。

然后调用GEMM（矩阵乘矩阵）库加速两矩阵相乘也就完成了卷积计算。由于按照计算需求排布了数据顺序，每次计算过程中总是能够依次访问特征图数据，极大地提高了计算卷积的速度。 （不光有GEMM，还有FFt（快速傅氏变换））