各位教授，大家早上好。

今天我要做的论文报告，题目为A High Performance Framework for Large-scale 2D

Convolution Operation on FPGA。

这一页是我的PPT的目录。总共会讲5个方面。第一部分是二维卷积的简单介绍，第二部分是分析原有工作的不足以及我们工作的动机。第三部分是对基本单元的改进，包括原有工作的基本单元设计以及我们改进的基本单元设计。第四部分是对于大规模卷积运算所提出的结构，包括对于片内存储的使用优化以及分块运算。第五部分是对结果的分析。

首先，我们对二维卷积运算做一个简单的介绍。The 2D convolution operation is to convolve the feature maps of the previous layer with a plurality of learnable kernels to form the feature maps of the next layer. 这里有一个比较生动的例子，图中大图片为输入特征图，黄色区域为运算窗口，下标为卷积核参数。每次运算将运算窗口内的输入特征图与卷积核进行对应位置的乘法，再累加所有的乘法结果，得到最终的运算结果。不断移动运算窗口进行计算，得到最终结果。 所有输入特征图与一组卷积核参数可以计算得到一副输出特征图。所有卷积核参数计算完毕，得到所有输出特征图。

可以得知，卷积计算既是运算密集型的，又是内存密集型的。大型卷积神经网络的乘法次数可以达到10 billion.

以前的工作的不足之处在于，首先，卷积运算基本单元的设计存在缺陷，当卷积运算的步长大于1时，也就是窗口每次滑动的步长大于1时，会存在大量的冗余计算。其次，大多数的工作都聚焦在小型卷积神经网络上，而由于FPGA片内存储资源的限制，这些结构无法应用于大型卷积运算。

我们的出发点在于，首先对基本单元进行改进，使得在步长大于1时，不会出现冗余的计算。然后，提出一种能够适用于大型卷积运算的结构，使得大型卷积运算可以在小型FPGA上进行加速，同时尽量减少片外访存带宽。使大型卷积神经网络可以付诸应用。

这一页讲的是传统的基本单元设计，叫做Z型基本单元。