各位老师大家好，我今天展示的毕业设计题目是《面向深度学习算法的异构加速计算技术研究》

（研究背景）

异构计算的意义在于，使用多种不同架构的硬件对同一计算任务进行处理，以达到最大程度地发挥各种架构的优势的目的。

神经网络算法常见的结构通常为多个神经元层的传播结构，同一层内神经元所完成的运算具有相互独立性，一般不相互依赖，这意味着神经网络模型天然具有很高的并行计算潜力。另一方面，神经网络中许多运算操作都可以转换为矩阵乘法，这意味着神经网络也可以使用GPU这类常用于向量运算和矩阵运算的器件进行加速。

（LeNet的部署）

本设计分别将LeNet部署在了CPU、GPU和FPGA上，主要使用的框架是pytorch和xilinx vitis提供的基于C++的HLS语法。

LeNet的结构如下，图片数据通过两个大卷积层，再通过两个全连接层和一个高斯连接层完成运算，最后生成识别结果。两个大卷积层又分别可以分解成一个卷积操作、一个Relu层和一个降采样层（池化层）。卷积操作经过一定的变换可以转化成矩阵乘法，全连接层的乘加操作也可以转化成矩阵乘法。另外，Relu函数实际上就是一个截断在原点处的一次函数，这意味着硬件上也比较容易实现这一激活函数。

LeNet在CPU和GPU上的部署较为相似，图中主要展示的是它们的训练过程。……（见ppt）

LeNet在FPGA上的部署与前两者差异较大。首先，作为硬件，FPGA是不宜完成参数优化训练的，因为参数的优化需要取得所有层的中间输出，以获得梯度分析。因此FPGA版本的LeNet要取用CPU或GPU版本训练好的参数。然后FPGA版本的LeNet运算结构如图所示，在CPU中运行的host程序先将图像数据和参数数据存放到内存中，然后搬移到FPGA的内存中，FPGA中部署的内核按照图示数据流进行运算，最终得到结果并返还到CPU的内存中，最终输出结果。

（前端设计）

本设计还提供了一个对LeNet程序状态的监测前端页面。这个页面的主要框架关系如图所示，……

前端页面要与LeNet程序进行交互，是一个较长的响应链。直接定位到内存空间并进行信息传输的方法在windows操作系统中行不通，因为windows操作系统会给不同的进程划分私有的内存空间，并且会通过碎片化处理的方式将其分页打断，以保证内存空间的安全性。因而本设计中网页端与部署程序端双方通过cache文件进行信息的传递。进程间通信的方案应该是可行的，受限于时间，这一功能还没有实现。

相应地，部署程序需要进行一定的修改。我们让部署程序默认进入idle状态，并通过轮询的方式接收指令，进入其他状态，执行完相应状态的任务后再回到idle状态。