**2020年新工科联盟-Xilinx暑期学校团队项目设计文档**

**设计文稿提交格式**

**(Project Paper Submission Template)**

|  |  |
| --- | --- |
| **作品名称** | 海量指纹库匹配任务加速机制设计 |
| **板卡型号** | PYNQ-Z2 |
| **所在班级** | B-14班 |
| **成员姓名、学号、学校** | 孙舒婷 20BS184A 重庆大学  时辰 20BS184B 重庆大学  徐丹丹 20BS184C 重庆大学  宁坤 20BS184D 重庆大学  仇实 20BS184E 重庆大学 |
| **Github链接** | *<https://github.com/Shootingholic/Finger-Match-Acceleration-Method.git>* |

**第一部分**

设计概述 /Design Introduction

（1.请概括地描述一下你的设计，可包括本设计目的、学习到的知识点、应用方向或者设想的应用场景等；2. 经组内成员讨论后以表格的形式描述项目中各成员在项目中发挥的作用或者贡献百分比；3.作品的展示照片）

# 1.设计概述

## 1.1设计目的

指纹由于其个体差异性及稳定性被广泛用于身份确认领域。现代指纹识别主要包括指纹图像的采样、预处理、分类、特征信息和匹配等过程。在前期的调研中，发现对指纹数据的预处理操作的执行效率对整个指纹识别的准确度、执行时延有至关重要的影响。指纹方向场作为数据预处理的关键步骤，其基本是基于神经网络实现的，而神经网络的推断在CPU平台上串行计算十分的低效，利用FPGA的神经网络可以更好的实现网络的并行计算与资源复用。因此，本项目针对指纹识别的方向场识别实现硬件加速设计，从而加速在海量指纹库中的指纹匹配速度。

## 1.2学习到的知识点

在本次项目的开展和实施的过程中，我们团队所学习到的知识主要分为四个部分：了解指纹识别算法、IP核的设计与优化、硬件平台的架构设计以及整体部署。

**指纹识别算法**

现代指纹识别主要包括指纹图像的采样、预处理、分类、特征信息和匹配等过程，其性能评价标准是系统识别率和执行速度。在实际应用中，所采集到的指纹图像都包含各种各样的噪声，这对识别的效果提出了挑战。幸运的是，在学术界对指纹图像方向提取、指纹图像增强、奇异点检测、指纹图像分类和细节点匹配等能保证指纹识别技术可靠性和鲁棒性的操作都有广泛的研究。其中，指纹图像方向场描述了指纹脊线的方向纹理信息，方向场是指纹图像稳定、可靠的先验知识，它在指纹图像增强、指纹奇异点检测、指纹图像分类及匹配等方面起着重要的作用。指纹方向场提取主要有：基于梯度、基于模型、基于滤波、基于卷积神经网络等类型。目前基于卷积神经网络的方法取得了很多研究成果，并且在指纹处于低质量的情况下提取效果较好。

**IP核的设计与优化**

通过对指纹识别算法的理论上的学习，我们参考FingerNet的网络结构，从中学习指纹方向场提取的算法实现。首先，参照FingerNet的指纹方向场提取部分，在HLS上实现了卷积神经网络关键层的函数：卷积层、池化层、批归一化、PReLU等，并写了相应的testbench验证每个函数的功能正确性。在这过程中，由于参考的网络结构是基于TensorFlow框架搭建使用python语言搭建的，在一开始我们对python语言不是很了解，网络还调用了很多框架中的函数，学习该网络对于我们来说是一个很大的挑战，也耗费了不少的时间，但幸运的是，在设计IP核的过程中，我们对Python语法、框架的关键函数功能有了一定的了解；再另一方面，还HLS与C/C++的语法兼容性有更深的了解，有些C/C++支持而HLS不支持的语法或者是因为HLS可支持定点数而定点数的乘法会导致结果溢出等问题，导致我们在设计IP核过程中频繁受阻，这也为后续的设计开发排雷。

在整合了卷积层、批归一化层、PReLu层后，发现资源占用十分大，严重超出PYNQ-z2所提供的片上资源，尤其是BRAM的资源。我们主要从三个方面来着手设计优化方案：

* 从函数本身出发，找到可融合的操作，避免冗余计算开销；
* 使用定点数，减少资源占用量
* 使用HLS提供Directive功能进行优化设计

在这优化过程中，最大的阻碍在于没有对HLS所提供的每个Directive的操作有一个直观的理解。通过不断的去试验，我们找出了比较合理的优化方案，这过程中让我们对Directive中的各操作有了多样的了解。

**硬件平台的架构设计**

该部分主要是PS与PL端的通信设计，具体来说就是利用vivado软件的block design模块将每个IP核布局布线。在这部分我们最主要的做的就是明确PS和PL的各部分的功能，然后针对其功能来设计数据通路。在前述的IP核设计我们就只考虑了算法的功能模块，而利用这些功能块还需要从片外读入权重、输入数据等在BRAM上做操作才能做后续的操作，因此针对这个我们还设计了一个专用的数据读入调度模块；在另一方面，在做了计算之后，数据应该存在哪也设计了专用的数据输出模块。在这部分，考验我们对全局的把控，我们从中主要学习到了在FPGA板上如何实现从内存中读数据，以及将如何将数据写入内存中。

**整体部署部分**

在整体部署时，我们遇上了片上资源不足的关键问题。而我们实现的网络是参照现有的广泛使用的网络FingerNet做的。在项目初期估计其所需资源时我们使用的PYNQ-z2板子能够满足，但在真正部署后，发现片上资源无法满足权重数据及输入数据的存放。常规来说，我们可以通过网络的剪枝、轻量化等来解决这个问题，而项目周期短，我们只能另寻蹊径，采用了两种改进的办法：一是，利用片上内存与片外内存的通信来弥补片上内存不足的缺陷，将上述数据放置在DRAM上；二是，从数据源本身出发，我们将输入的指纹数据降低分辨率同时做切割，以此来减少数据存储开销。然而，我们的改进方法在很大程度上限制了样本的多样性也引入了时延。在这之中，我们学习到了再做硬件部署前，对资源分析的重要性。

由于计算所需的数据权重大，只使用一个接口来访问DRAM存在数据争用端口而造成数据传输时延长，在寻求解决方案时，我们尝试了无数的方式，最后采取了对DRAM地址空间设置多个端口的形式来加快数据传输。

## 1.3设想的应用场景

对于指纹数据，我们最常想到的就是公安利用指纹破案的场面，而在犯罪现场采集到的数据一般都有不清晰、噪声数据很多的缺点，利用这种数据在海量的指纹库里面去坐对比来找到犯罪嫌疑人，会存在两个问题：第一，采集数据不清楚；第二，指纹数据库是分布存储,根据不同地域,将指纹存储在不同地方，每次都要将边端指纹数据传到云端去做匹配，十分的耗时。因此，我们就想将该加速网络部署在边端指纹库上，加速在海量指纹库中与污损指纹匹配的速度。

# 2.成员贡献

|  |  |  |  |
| --- | --- | --- | --- |
| 姓名 | 学号 | 项目中发挥的作用 | 贡献百分比 |
| 孙舒婷 | 20BS184A | 成员任务分工，IP核的设计，整体架构的搭建，GitHub文件上传和维护 | 25 |
| 时辰 | 20BS184B | FingerNet方向场提取网络的分解，IP核的设计及优化，整体架构的搭建 | 25 |
| 徐丹丹 | 20BS184C | IP核的设计及优化，文档的书写与整理 | 15 |
| 宁坤 | 20BS184D | IP核的设计及优化，空洞卷积的实现 | 10 |
| 仇实 | 20BS184E | IP核的设计及优化，解决资源不足的问题 | 25 |

# 3.作品照片展示

**第二部分**

系统组成及功能说明 /System Construction & Function Description

（请对作品的1. 计划实现及已实现的功能；2. 项目系统框图；3. 使用的技术方向做说明）

# 计划实现及已实现的功能

## 1.1计划实现

计划实现的内容如图一所示，系统架构分为PS与PL两个部分，我们在PL部分实现方向场提取硬件加速器。利用DMA通过AXI4-Lite和AXI4-Stream读取片下的DDR中指纹图片数据到buffer中，加速器与片上的buffer进行数据交互并对指纹图片进行方向场提取，向用户输出方向场提取效果。

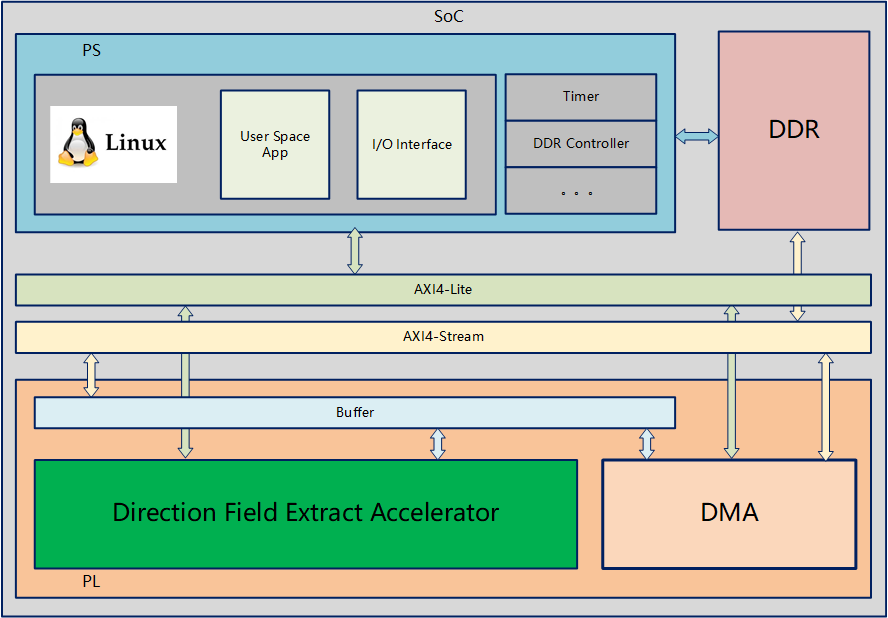


图 1 系统硬件架构图

## 1.2已实现

* 以[CNNIOT](https://github.com/mfarhadi/CNNIOT)为设计参考，我们构建了可以通过发送不同数据包头来实现不同网络层计算的cnn加速IP核。
* 对IP核进行了轻量化操作，根据软件计算结果中的精度表示以及结果验证，我们发现小数点后6位的值对于模型的影响不大，因此可以用定点小数来进行网络模型轻量化，减轻片上BRAM的压力。
* 以[Anguli](https://dsl.cds.iisc.ac.in/projects/Anguli/index.html" \l "about)提供的指纹生成器所产生的指纹数据集，成功运行了方向场提取部分的网络，但是还是由于片上资源有限，输入图像需要实施非常大的压缩率才能将网络完成，因此在指纹方向场的提取效果上并不如预期。
* 完善了驱动文件的编写，向原加速器驱动文件中添加了空洞卷积和批量归一化层调用的驱动接口。
* 编写了方便查看方向场提取结果的可视化软件程序。

# 2.项目系统框图

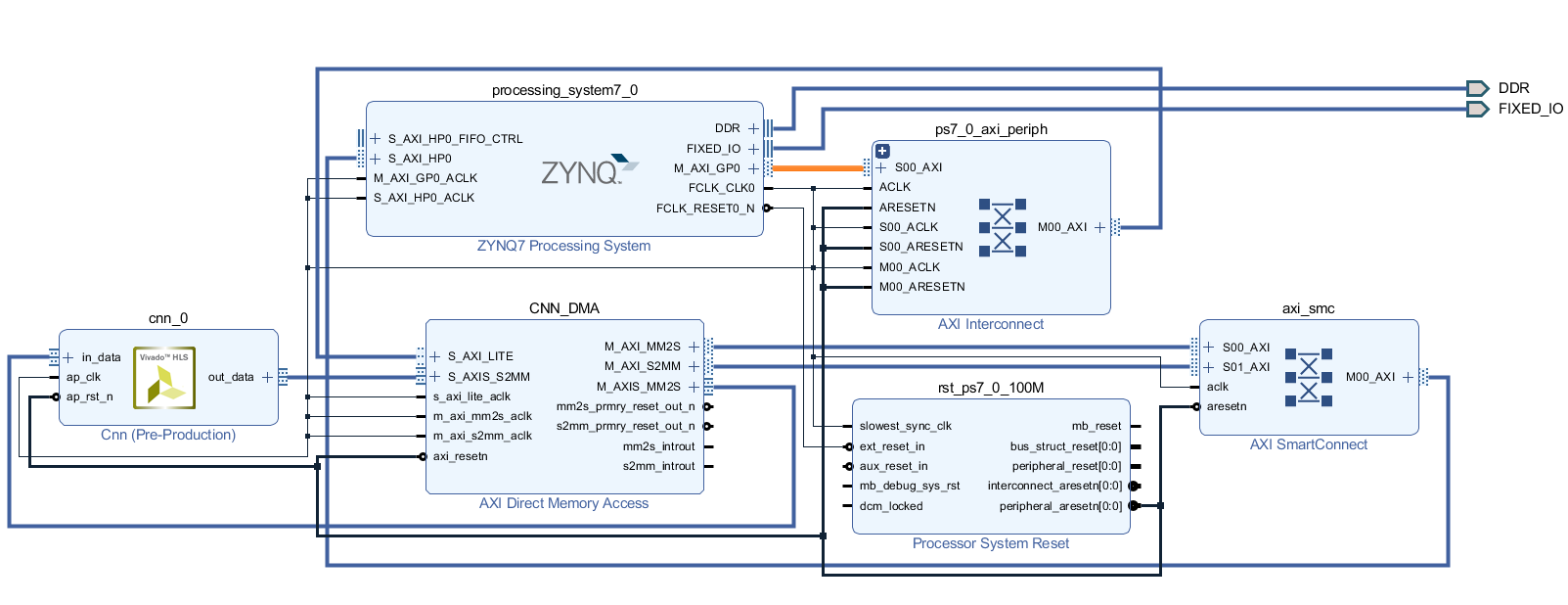


图 2 系统通路设计图

# 3.使用的技术方向做说明

针对海量指纹库匹配任务加速机制这个项目的实现上，我们团队所做的工作主要可以分为三个部分。第一，IP核的设计与优化；第二，系统架构的设计；第三，针对FPGA资源受限，如何将算法功能部署在FPGA上。

**IP核的设计与优化**，IP核的设计分类，一类是功能IP核设计，一类是输入输出IP核的设计。功能IP核设计是指利用IP核去实现指纹方向场提取的算法。利用指纹方向场提取的理论说明完成C/C++代码的编码挑战难度较大，于是我们采用的是将常用的指纹识别网络FingerNet分解，分解出方向场识别部分，并验证分解出来的网络子模块的准确性。分解出的FingerNet的子模块是基于TensorFlow框架的Python语言实现，不能将其直接使用。在该部分我们先对FingerNet子模块进行Python—>C/C++的转换，再根据综合出来的报告对其做优化，优化策略有

* 从函数本身出发，找到可融合的操作，避免冗余计算开销；
* 使用定点数，减少资源占用量；
* 使用HLS提供Directive功能进行优化设计。

优化的目标为，减少资源用量，使计算所需要的资源能满足片上资源限制，同时尽量的降低时延。

本项目中将实现方向场提取的卷积神经网络，在计算优化的结果之上，我们还从数据的读取和存储方面，对IP核进行了优化设计，基于卷积层的从片上内存（BRAM）或者DRAM的中读取权重数据、输入图像数据等的需求，结合Zynq芯片特性和AXI总线进行了数据接口设计。在产生了输出之后，也需要将神经网络的输出结果存到相应的位置。结合了计算功能和数据搬移功能的IP核即为这一部分工作的交付物。

**系统架构的设计，**系统架构设计主要内容设计合理的数据通路以保证网络运行结果的正确性。这部分设计主要包含两个部分，一是通过Zynq芯片上的ARM核协同片上的快速总线axi，实现应用数据及结果的搬运，实现完善的数据流，其控制代码可以在基于Overlay、lasagne、pynq等库的基础上，利用Python语言实现；另一部分的主要目的是保证网络中间数据在各层之间的正确流动，通过复用可配置的各类网络层的计算IP核，构建本项目中计划实现的FingerNet中的方向场提取的卷积神经网络部分，该部分可配置网络层的IP核的具体实现方式为设计统一长度的通信报文，向硬件提供配置信息。

**针对FPGA资源受限，将算法功能及模型数据在FPGA上完成部署。**前述的IP核的设计与优化主要是对单个IP核去考虑，但将整个网络部署在硬件上依旧有硬件资源不够的问题，需要在宏观上来使方向场提取网络成功部署在FPGA上。首先，我们面临的是计算所必须的权重数据和输入数据无法放在BRAM上，甚至单单说权重数据也无法放在BRAM上，因此我们考虑并尝试了两种改进的办法：一是，利用片上内存与片外内存的通信来弥补片上内存不足的缺陷，将上述数据放置在DRAM上，并且将DRAM分块，利用多端口访问，加快访存速度，此种方式会因为数据局部性遭到破坏，应该在缓存中取得的数据却每次都要从DDR中读取，降低了数据重用性的后果必然会导致性能的断崖式下降，这与加速目标是十分相悖的；二是，从数据源本身出发，我们将输入的指纹数据降低分辨率同时做切割，以此来减少片上BRAM的数据存储开销，此外，也需要使用定点数量化的方式，减少片上内存的需求，我们通过反复的核算和实验，才在资源非常有限的Pynq上，实现了整个网络的部署，虽然以一定的精确度损失为代价，但相比上一个方法已经具备一定优势。

**第三部分**

完成情况及性能参数 /Final Design & Performance Parameters

（作品已实现的功能及性能指标）

主要功能的代码都已完成，但是还在加班加点debug。目前没有可视化的结果可以输出，争取今天之内能解决。

**第四部分**

总结 /Conclusions

（谈一谈完成暑期学校课程后的收获与感想。请每位组员分开写。）

**孙舒婷**

本次项目中，我通过在课堂上对xilinx软硬件开发流程的学习，越发深刻的了解了Zynq芯片架构下的工作机制，并且通过小组项目实践，更进一步地理解和学习pynq框架下的软硬件编程方式，在硬件IP核的优化设计中，感觉对多种不同编译指令的功能理解还稍有欠缺。在组员们共同遇到问题解决问题，产生问题消灭问题的过程中，加深了我们小组的凝聚力，增强了我们的行动力。我作为组长，在整个开发过程的任务分配和人员调度方面，感觉还需再锻炼，不然我们应该能提交更完善的作品。

**时辰**

本次的暑期学习使我收获颇丰，期间不仅通过课程学习和线上讲座补充了很多我之前没有接触过的知识，包括如何使用PE阵列做卷积运算，还通过课程作业和项目开发过程丰富了对vivado hls工具和pynq的认识，也对pynq产生了很浓厚的学习兴趣。在学习过程中，也遇到了很多难题，比如自己一直缺乏对于数据接口配置方面的经验，对于一些directive的使用还不太熟练等等，还需要通过不断的接触来补充这方面的知识。

**徐丹丹**

通过此次的暑期学校课程讲解及项目设计，我的收获远远大于预期。在课程讲解中，不仅学习到了硬件加速的开发流程，还拓宽了视野，增加FPGA以及Xilinx配套产品的知识。在项目设计实战中，将课上所学到的内容实际应用上，还是遇到了很多问题。首先，对HLS开发工具没有十分熟悉，总踩坑出错耗时间，所幸的是还是找到了问题所在，以后也能避开了本次所涉及的坑；其次，在项目设计之前，从未做过利用FPGA实现完整的计算开发相关的工作，通过和小组成员配合，了解了数据通路设计的相关工作，并且在问题解决上也学到了不少的技巧；最大的收获还是在于对FPGA的整个开发流程有过了实操，不再是理论上的概念，让我很有实感；最后，我对这方面的熟悉程度远远不足于队友，很感谢队友们能耐心的解答我的问题。

**宁坤**

首先很感谢也很庆幸能参加这次xilinx的暑期培训项目。参加这次项目是由于本身对使用HLS工具在FPGA上开发硬件加速器有一定的了解和兴趣，然而并没有很完整的参加一次项目的开发，想通过这次暑期的项目经历一次开发的全过程。在本次项目中，通过老师的授课，我对FPGA这种可编程逻辑阵列内部结构有了更深层次的了解，了解了HLS内部的工作机制，学习到了HLS的三大类优化参数的使用及一般常用的优化思路。此外，对pynq架构有了一定的了解，也认识了vitis AI的架构，同样也使自己对这两个架构产生了浓厚的兴趣，产生了继续深入学习的想法。最后经过完整的开发项目的过程，对HLS优化有了更深层的了解，也掌握了接口设计的基本知识，以及如何使用DMA在cpu和fpga中收发数据，也锻炼了自己独立开发的能力。

**仇实**

本次暑期学校的学习和实践中，我在编程和调试程序方面得到了可观的能力提升，原本在理论方面自己进行学习的时候，觉得所知甚浅，“绝知此事要躬行”，在实践中，对于硬件加速、系统设计、存储管理等方面，都有了非常长足的理解。虽然我们的项目成果距离预期仍有一段距离，但是这个项目同时也激发了我们进一步进行深究的兴趣，与小组成员的密切合作也为我们往后的科研工作打下良好基础。