**分 类 号 学号 M202072547**

**学校代码 1 0 4 8 7 密级**



**硕士学位论文**

**二值虚数神经网络(BCNN)专用加速芯片的研究与设计**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **学位申请人** | **：** |  | **高余敬** |  |  |
| **学科专业** | **：** |  | **电子信息** |  |  |
| **指导教师** | **：** |  | **童乔凌 教授** |  |  |
| **答辩日期** | **：** |  | **XXXX年XX月XX日** |  |  |

**A Dissertation Submitted in Partial Fulfillment of the Requirements**

**for the Professional Master Degree**

**[Research](https://www.bing.com/dict/search?q=research&FORM=BDVSP6&cc=cn)****[and](https://www.bing.com/dict/search?q=and&FORM=BDVSP6&cc=cn)****[Design](https://www.bing.com/dict/search?q=design&FORM=BDVSP6&cc=cn) of Bin[ary](https://www.bing.com/dict/search?q=Binary&FORM=BDVSP6&cc=cn)****[Complex](https://www.bing.com/dict/search?q=complex&FORM=BDVSP6&cc=cn)****[Neural](https://www.bing.com/dict/search?q=neural&FORM=BDVSP6&cc=cn)****[Networks](https://www.bing.com/dict/search?q=networks&FORM=BDVSP6&cc=cn)****[(](https://www.bing.com/dict/search?q=%28&FORM=BDVSP6&cc=cn)BCNN****[)](https://www.bing.com/dict/search?q=%29&FORM=BDVSP6&cc=cn)****[Inference](https://www.bing.com/dict/search?q=inference&FORM=BDVSP6&cc=cn) Chip**

**Candidate : Gao Yujing**

**Major : [Electronic](https://www.bing.com/dict/search?q=electronic&FORM=BDVSP6&cc=cn) [Information](https://www.bing.com/dict/search?q=information&FORM=BDVSP6&cc=cn)**

**Supervisor : Prof. Tong Qiaoling**

**Huazhong University of Science and Technology**

**Wuhan 430074, P. R. China**

**October, 2022**

独创性声明

本人声明所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除文中已经标明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本文的研究做出贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：

日期： 年 月 日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，即：学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权华中科技大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

保 密□，在 年解密后适用本授权书。

本论文属于

不保密□。

（请在以上方框内打“√”）

学位论文作者签名： 指导教师签名：

日期： 年 月 日 日期： 年 月 日

# 摘 要

机

**关键词：**

# Abstract

As

**Key words:**

目 录

二值虚数神经网络(BCNN)专用加速芯片的研究与设计 I

A Dissertation Submitted in Partial Fulfillment of the Requirements II

for the Professional Master Degree II

摘 要 I

Abstract II

1 绪论 1

1.1 研究背景与意义 1

1.2 研究现状 2

1.3 本文主要研究内容 9

1.4 本文的章节安排 10

2 CNN与BCNN算法概述 12

2.1 引言 12

2.2 卷积神经网络原理 12

2.3 BCNN算法概述 12

2.4 Zero\_DCE++算法概述 12

2.5 本章小结 12

3 面向BCNN算法的协处理器设计 13

3.1 引言 13

3.2 协处理器设计目标与架构 13

3.3 数据通路及其并行性设计 13

3.4 内存设计及数据存储方式 13

3.5 脉动阵列设计 13

3.6 Batch Normalization硬件设计 13

3.7 本章小结 14

4 BCNN专用加速器的系统设计 15

4.1 引言 15

4.2 RISC-V处理器及其协处理器 15

4.3 指令集（ISA）设计 15

4.4 系统硬件架构设计 15

4.5 系统软件架构设计 15

4.6 本章小结 15

5 仿真与测试 16

5.1 引言 16

5.2 功能仿真 16

5.3 资源评估 16

5.4 本章小结 16

6 总结与展望 17

6.1 本文主要内容及结论 17

6.2 本文的主要创新点 17

6.3 展望 17

致 谢 18

参考文献 19

附录1 攻读硕士学位期间取得的学术成果 20

附录2 部分程序代码 21

# 绪论

## 研究背景与意义

近年来，随着人工智能和深度学习的研究的不断深入，各类人工神经网络在各个应用方面都取得了丰硕的成果，特别是深度神经网络（DNNs）在许多计算机视觉应用都取得了极大的成功，如图像分类、目标检测等。而为了成功将神经网路应用推向资源受限的边缘设备，研究人员也一直在不断调整DNN模型。在这一过程中二值神经网络（BNNs）凭借其单比特神经元特性在众多的模型中脱颖而出。与其他通过裁剪神经元个数不同，二值神经网络通过将每个神经元的比特数压缩到极限来满足受限的硬件开销，即BNN每层的输入、权重和激活层的每个元素都仅使用单个二值化的数值（0或者1，分别代表-1或+1）来表示。

二值神经网络对于嵌入式应用具有如下的几个优势：（1）计算效率：二值网络具有极强的硬件友好性，通过将原先的32或者64比特的神经元映射到单比特，而原先的全精度运行也相应的映射为单比特的异或运算(xor)外加一个计“1”（pop-count）操作，将计算效率提升了不止10倍。从内存角度来说，使用单比特替代32比特的单精度浮点数或16比特的半精度浮点的使用，二值神经网络更是大幅度提升了内存的利用率和带宽。（2）低硬件开销：由于硬件逻辑结构简单和极低的内存需求，二值神经网络的硬件开销远远小于其他的神经网络。（3）能效比：因为其低硬件开销和功耗低等特性，基于二值网络的硬件对移动设备的电池也是十分友好的。由于这些优势，二值神经网络已经被广泛运用，如自动驾驶、智慧农业、COVID面部识别、图像增强等。

虽然二值神经网络具有以上有点，但是由于二值化过程中信息的丢失和模型的压缩，导致BNNs的精度损失也很严重。所以自从二值神经网络提出后，大量的研究人员也在致力于提高该网络的精度。二值复数神经网络便是其中一项重要的贡献。

复数神经网络的提出本意在于修正普通的DNNs网络，因为大多数DNNs网络都是采用实数来表示输入和权重，与复数网络（输入、输出和权重都是用复数来表示）相比，不论是表示能力、泛化能力还是抗噪声能力都略逊一筹。大量的实验结果也证明了在相同大小的网络中，复数网络具有更高的准确性。且另一个引人注目的优势便在于复数网络可以携带相位信息，这是其他网络没有的。

而集成了二值神经网络和复数神经网络的二值复数神经网络既继承了BNN网络的计算效率、低硬件开销、高能效比等优势的同时，又结合DCN网络提高了BNN网络的精度。在二值复数神经网络中，每层的输入、权重和输出都使用二值化的复数值来表示，{1+i,1-i,-1+i,-1-i}中某一个，即每个神经元使用双比特来表示（一个表示实数部分，另一个表示虚数部分）。而二值复数卷积操作依旧可以使用xor-popcount运算替代了大量的乘加计算，具有极强的硬件友好性。

另一方面，随着人工智能和物联网时代的到来，如何将神经网络部署到边缘终端应用也成为了另一项重要课题。由于卷积神经网络规模的不断扩大，现有的计算机处理器（CPU）受限于“存储墙”的影响，越来越无法满足当下神经网络对数据处理的要求，而使用GPU虽然可以满足神经网络的计算需求，但是其功耗和成本都让其难以在推广到边缘设备，所以针对不同的神经网络及应用场景来设计专用的加速器是十分有必要的。

加速器可以在机器学习的各个阶段得到应用，无论是训练还是推断，边缘端设备或者是云端设备。不同的应用场景对加速器也有着不同的要求，譬如延迟、能耗、吞吐量、面积、可编程能力和系统集成。

如果没有专用的加速器，如何高性能低功耗地起去运行一个先进准确的DNN网络是极其困难的问题。目前为止，虽然二值复数神经网络的算法层面的研究已经逐步展开，而针对二值卷积神经网络的加速器的研究工作还没有完全开展，基于此种情况下，我们开展了面向二值复数神经网络的专用加速器的研究与设计工作。

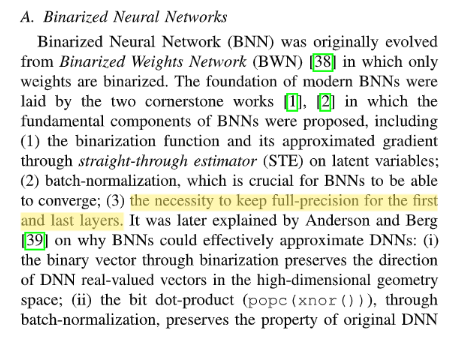
## 研究现状

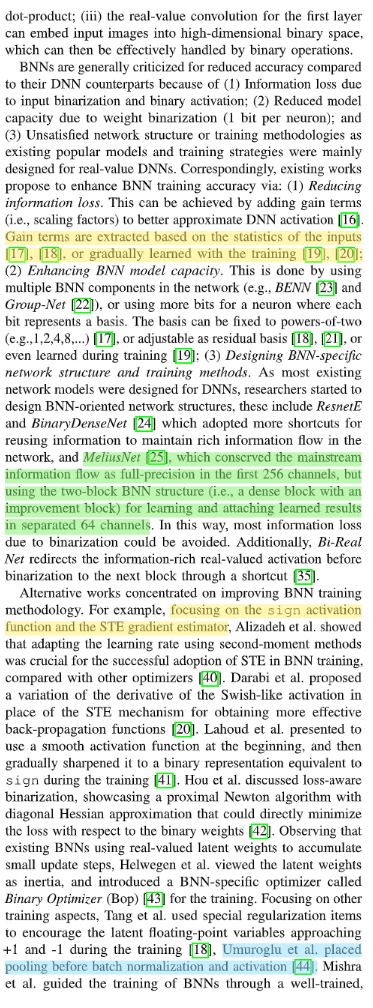
### 二值复数神经网络的研究现状

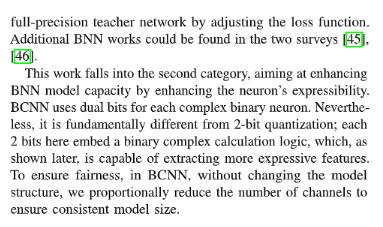
二值复数神经网络的发展时间较短，由浙江大学的Yanfei Li等人在2021年首次提出，所以面向该网络的研究成果还不是很多。而二值复数神经网络紧密联系了二值神经网络和复数神经网络，故BCNNs的研究现状与二者是密不可分的，可以看作是二者拓展并交叉的新一领域，故讨论BCNN的研究现状自然也可等价为研究BNNs和DCNs两个神经网络的现状。

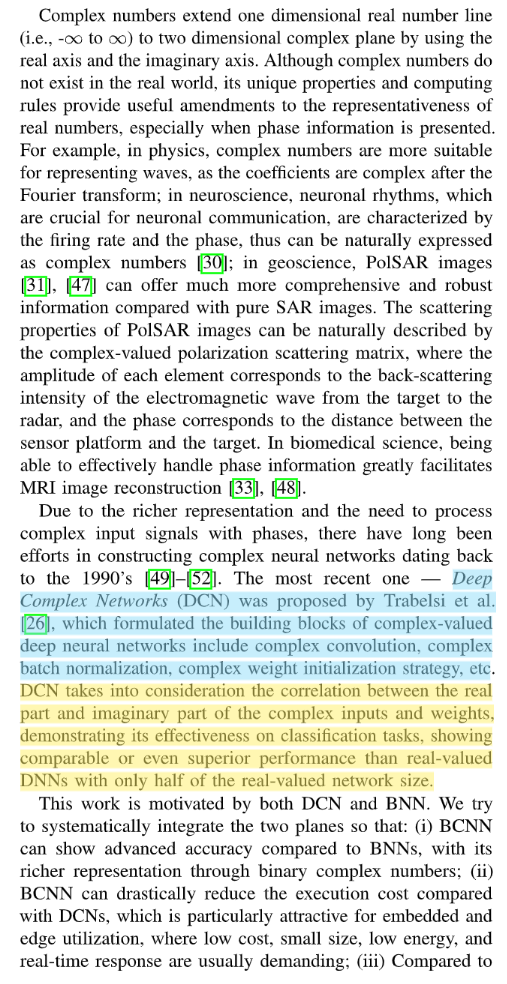
二值神经

M. Courbariaux I. Hubara

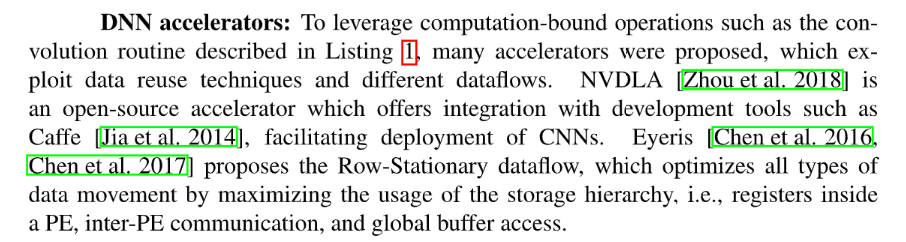


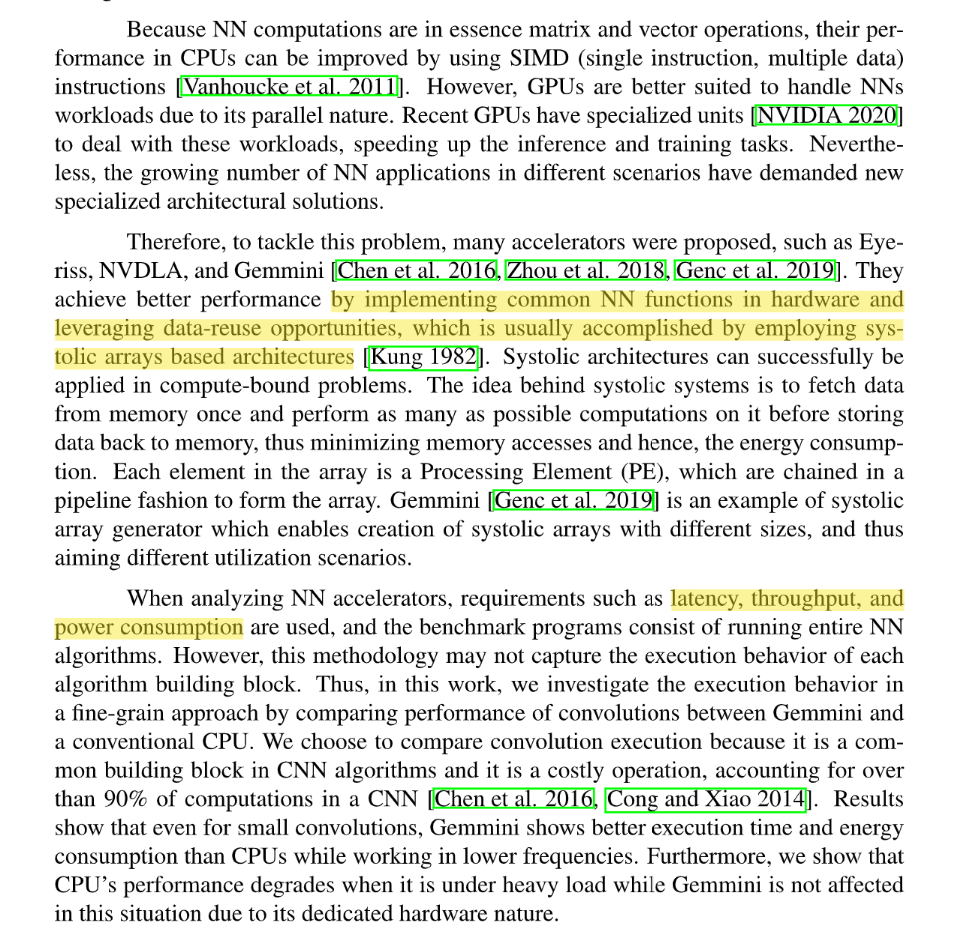


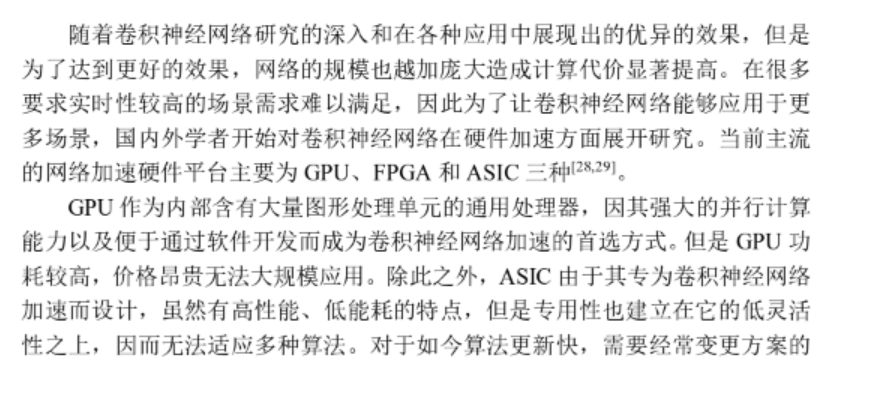


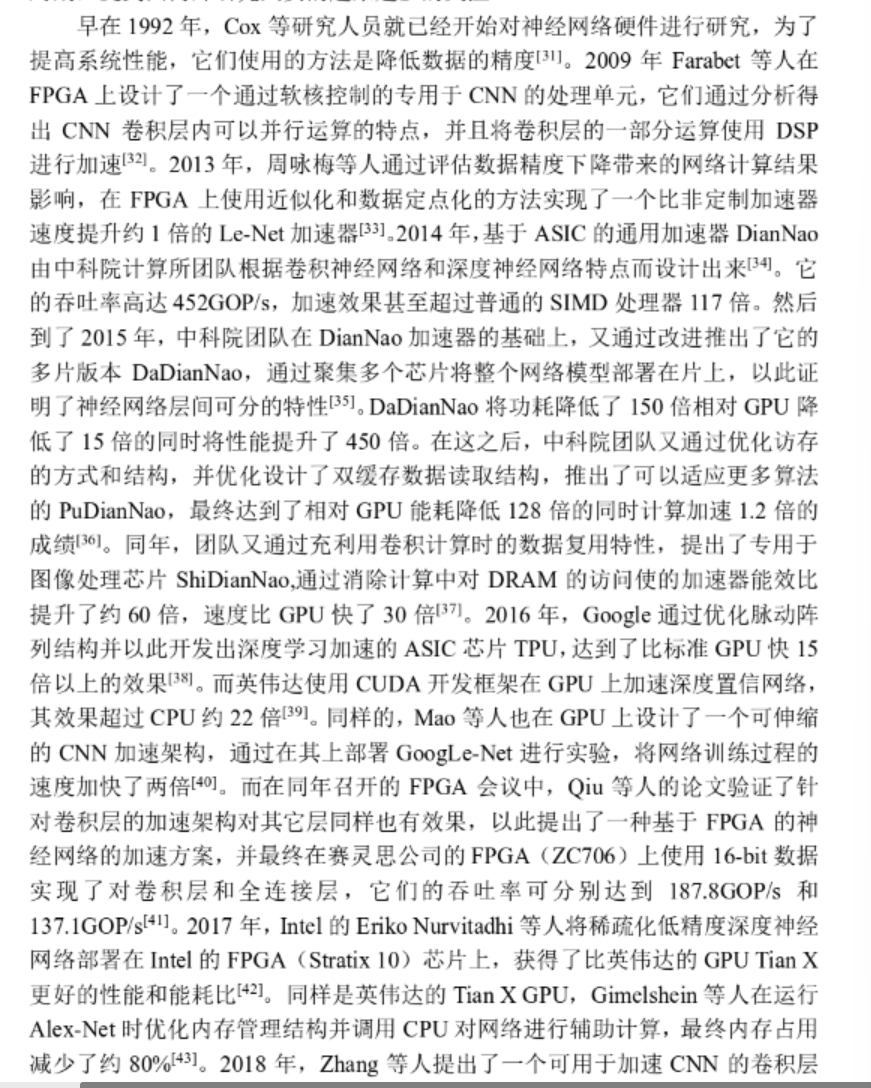


### 神经网络加速器的研究现状









## 本文主要研究内容

本文针对二值复数神经网络的专用加速器进行研究与设计，致力于解决深度卷积神经网络部署到边缘端设备时计算资源、存储资源占用过多等缺点，和二值神经网络部署后精度损失严重等问题。与传统的算法专用加速电路不同，本次设计的专用加速器，不仅包括了针对二值复数神经网络加速任务的协处理器，还将其集成到一块完整的SOC中（其中包含RISC-V核Rocket,DRAM以及数据Cache等），并在此基础上实现了专用于协处理器的扩展指令，实现了“全栈式”的软硬件协同。

后续章节首先从卷积神经网络层面进行分析，研究卷积神经网络的相关概念，并分析不同数据流的脉动阵列是如何从硬件角度实现卷积神经网络的，其中包括weight-stationary和output-stationay两种数据流。之后从二值复数神经网络的算法层面着手分析，研究算法本身的并行结构、数据特点，以及二值复数的计算、象限二值化、batch normalization等操作，其中首要在于分析二值复数神经网络的具体模型以及各层功能与具体实现，并基于“软硬件协同”设计理念分别提出硬件实现方案。

在有了上述算法研究的基础上，我们将设计出适用于各类二值卷积神经网络算法的协处理器。基于“软硬件协同”思想，我们首先提出了协处理器的指令集架构，但是为了更好的说明该指令集架构，我们在文章中先对硬件设计进行说明；在硬件层面我们首先设计协处理器的数据通路并讨论数据通路的并行性，其次是协处理器的内存与数据存储方式设计，紧接着计算核心脉动阵列的设计，其中包含了复数的卷积运算实现和xor-popcount运算实现等等。

完成上述的协处理器设计后，我们将协处理器与主处理器进行集成，并提出了面向二值复数神经网络的加速器的系统硬件整体架构，并针对完成的SOC系统进行软件设计，即开发相应的C函数库，在此函数库的基础上我们进而完成了二值复数神经网络在Zero\_DCE++应用的软件实现。

将上一步使用C语言完成的软件算法编译并在我们设计的专用加速器上运行，这一步也相当于完成了加速器的功能仿真验证。完成相应的功能仿真后，在FPGA板上实现加速器硬件电路，并进行资源评估。

## 本文的章节安排

本论文利用“软硬件协同”设计思想，设计了面向二值复数神经网络的专用加速器系统电路和底层函数库，在软硬件的基础上实现了Zero\_DCE++算法。

以下是本文的章节结构安排：

1. 绪论：本章主要通过阐述二值神经网络和复数神经网络的研究背景和研究现状来引入对二值复数神经网络的研究，并结合对边缘端加速设备的研究提出了本次课题的主要研究内容，最后介绍了本文的章节安排和主要内容。
2. 本章主要为卷积神经网络和二值复数神经网络的理论概述，我们首先阐述了卷积神经网络的基本原理和硬件实现方法，其次我们重点说明了二值复数神经网络的组成、二值化原理和二值复数计算原理，最后我们又介绍了Zero\_DCE++算法的应用。
3. 本章主要针对二值复数神经网络设计了专用的协处理器，我们首先提出了本次协处理器设计的目标与架构，其次对数据通路和内存及其相关细节进行说明与设计，然后又利用脉动阵列实现了二值复数的卷积运算，最后我们设计矩阵向量阈值模块来实现BCNNs网络的BN层和sign函数。
4. 在第三章协处理器设计的基础上，本章进而实现了二值复数神经网络的专用加速器的系统设计。为此，我们首先介绍了RISC-V开源环境，包括开源处理器Rocket和协处理器Gemmini等，之后基于RISC-V的扩展指令集提出了本次设计的指令集架构，在此基础上我们提出并实现了BCNN网络的专用加速器的系统设计。在指令集架构的基础上，我们又完成了对加速器的软件系统的设计并通过软件库实现Zero\_DCE++算法。
5. 本章主要是对前面实现的专用加速器进行功能仿真与性能评估。首先在RTL级的加速器上实现Zero\_DCE++算法，并与计算机运行结果对比，评定本次设计的专用加速器是否可以准确的完成二值复数神经网络的加速任务；之后在FPGA板上实现本次设计的专用加速器电路，并进行性能与资源评估。
6. 总结与展望

# CNN与BCNN算法概述

## 引言

## 卷积神经网络原理

### 卷积神经网络原理概述

### 脉动阵列

## BCNN算法概述

### 二值复数及其计算

### 象限二值化

### 复数Batch Normalization

## Zero\_DCE++算法概述

## 本章小结

本章主要针对

# 面向BCNN算法的协处理器设计

## 引言

## 协处理器设计目标与架构

### 设计目标

### 设计架构

## 数据通路及其并行性设计

### 数据通路设计

### 并行性设计

## 内存设计及数据存储方式

### 内存设计

### 数据存储方式

## 脉动阵列设计

### Weight-stationary型脉动阵列

### 复数卷积运算单元设计

## Batch Normalization硬件设计

## 本章小结

# BCNN专用加速器的系统设计

## 引言

## RISC-V处理器及其协处理器

### 开源处理器Rocket概述

### 协处理器Gemmini概述

## 指令集（ISA）设计

### 数据移动指令

### 配置指令

### 配置指令

## 系统硬件架构设计

## 系统软件架构设计

## 本章小结

本章

# 仿真与性能评估

## 引言

## 功能仿真

## 性能评估

## 本章小结

# 总结与展望

## 本文主要内容及结论

## 本文的主要创新点

## 展望

# 致 谢

# 参考文献

# 附录1 攻读硕士学位期间取得的学术成果

# 附录2 部分程序代码