東南大學

毕业设计(论文)报告

题 目： 生理信号检测卷积神经网络硬件加速器设计

学　　号：	06117113
姓　　名：	吴中行
学　　院：	电子科学与工程学院
专　　业：	物联网工程
指导教师：	刘昊
起止日期：	2020.12.1 至 2021.6.10

东南大学毕业（设计）论文独创性声明

本人声明所呈交的毕业（设计）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，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得东南大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

论文作者签名：_____ 日期：_____年_____月_____日

东南大学毕业（设计）论文使用授权声明

东南大学有权保留本人所送交毕业（设计）论文的复印件和电子文档，可以采用影印、缩印或其他复制手段保存论文。本人电子文档的内容和纸质论文的内容相一致。除在保密期内的保密论文外，允许论文被查阅和借阅，可以公布（包括刊登）论文的全部或部分内容。论文的公布（包括刊登）授权东南大学教务处办理。

论文作者签名：_____ 导师签名：_____  _____

日期：_____年_____月_____日 日期：_____年_____月_____日

摘　要

随着社会老龄化加剧和智能医疗领域的快速发展，患者对于便携式的智能诊断系统的需求也越来越大。传统的生理信号检测算法和智能诊断硬件设备在识别准确率、识别效率、便携性和低功耗方面的不足逐渐凸显。

针对这些问题，本课题将卷积神经网络与生理信号检测算法相结合，提出了一种基于卷积神经网络的新型生理信号检测算法，并针对网络模型提出了优化，极大地压缩了网络模型大小。同时，设计了算法对应的硬件加速器电路，针对访存功耗、存储功耗和计算功耗进行了优化设计。

本课题以 MIT-BIH 心律失常数据库为网络训练数据集，采用软硬件协同的开发方式，完成了基于卷积神经网络的新型生理信号检测算法及硬件加速器电路的设计。最终，算法模型识别准确率为94.7%，硬件加速器电路识别准确率达到89.7%，单次识别速度只有4μs，功耗仅有 0.651W。满足了预期的设计指标，基本满足了智能医疗诊断的应用需求，为生理信号检测提供了一种更科学、更具应用前景的解决方案。

关键词：生理信号检测算法，卷积神经网络，硬件加速器

ABSTRACT

With aging intensifying and the rapid development of intelligent medical field, patients have more and more demands for portable intelligent diagnosis system. Traditional physiological signal detection algorithms and intelligent diagnosis hardware devices have shortcomings in recognition accuracy, recognition efficiency, portability and low power consumption.

In order to solve these problems, this paper proposes a new physiological signal detection algorithm based on convolution neural network by combining convolution neural network with physiological signal detection algorithm, and optimizes the network model, which greatly reduces the size of the network model. In addition, the hardware accelerator circuit corresponding to the algorithm is designed, and the memory access power consumption, memory power consumption and computing power consumption are optimized.

The network training data set in this topic is MIT-BIH arrhythmia database. A new physiological signal detection algorithm based on convolutional neural network and hardware accelerator circuit are designed by using the software and hardware collaborative development method. In the end, the recognition accuracy of the new physiological signal detection algorithm is 94.7%, and the recognition accuracy of the designed hardware accelerator circuit is 89.7%. The single recognition speed is only 4μs, and the power consumption is only 0.651W W. It meets the expected design indexes, basically meets the application requirements of modern intelligent medical diagnosis, and provides a more scientific and promising solution for physiological signal detection.

KEY WORDS: Physiological signal detection algorithm, CNN, DLA

目　　录

第一章 绪论

1.1 研究背景及意义

中国医学科学院 2018 年的研究数据显示，心率失常是心血管病的主要病因，心律失常引发的心血管病是威胁人民健康生命的主要疾病。我国每年约有 54 万人死于心律失常，大部分猝死是医院外的延误治疗所导致的[1]。“早预防、早发现、早治疗”是避免这种现象发生的重要措施。

生理信号是检测一个人的健康状况的重要途径。其中，心电图(Electrocardiography, ECG)是检测心律失常的重要手段之一，在临床医学中广泛应用于诊断心律失常等心血管疾病[2]。有些类型的心律失常早期症状由于持续时间短，不易被察觉，但发病突然且症状强烈，如不及时治疗，可能导致中风、心力衰竭等后果，严重时可能导致死亡[3]。

如图 1-1 所示，诸如贴片式设备或手环等可穿戴设备可以长时间的实时检测心率，然后发送患者的健康数据给医生，便于记录、跟踪和诊断患者的心率变化情况。如果在患者出现心律失常等心血管疾病时，可以及时发出健康预警，避免中风、心力衰竭或猝死等更严重的情况发生。因此，在相关的医学领域，可携带式的心率检测设备已经得到了广泛关注[4]。

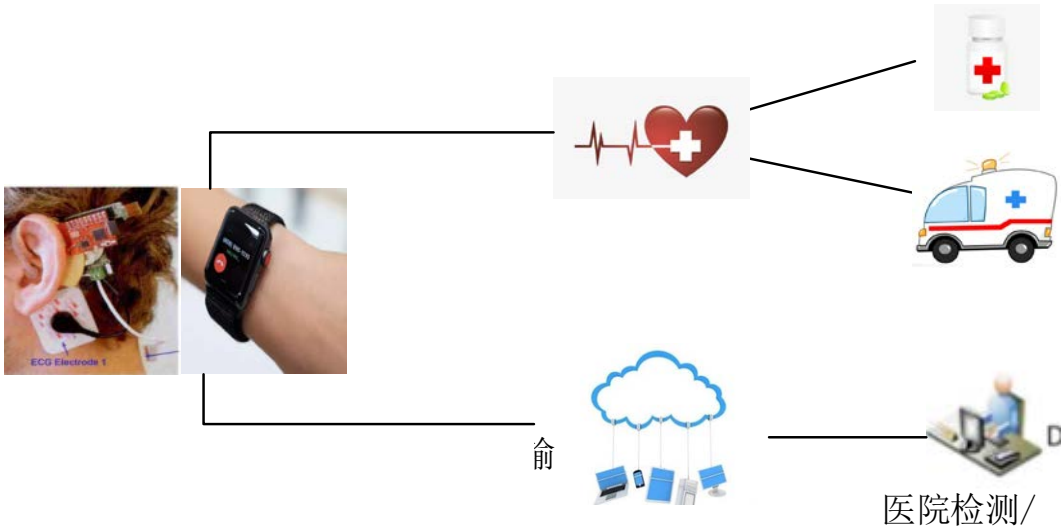


图 1-1　ECG 信号采集、检测及诊断过程

在当前医学研究中，对 QRS 复合波的分析比对长时间的心电图信号片段的分析更受欢

迎[5]。因此，现有的心律检测设备系统基本基于模式识别算法，而不同患者的 ECG 信号的 QRS 波存在很大差异，导致这类主要依赖于特征提取方法设计水平的检测系统泛化能力较低，无法很好地应用于数量庞大的心律失常患者群体[4]。

近年来，神经网络算法不断发展。相比特征提取的方法，神经网络算法具有从数据中学习更高级特征的能力，其层次化的结构使其拥有了更好的泛化能力和鲁棒性。神经网络直接从原始数据学习特征，不需要额外的数据处理和特征工程，更适合对长时间心电图信号片段的分析。因此，将神经网络(Neural network, NN)应用于智能心率检测设备的研究受到越来越多的关注[4]。

其中，卷积神经网络(Convolutional Neural network, CNN)最常用于处理二维数据，包括图像[6]等数据。由于心电图的周期性等数据特点，与其他深度学习架构相比，CNN 在处理心电图数据时突出了其良好的抽象能力和分类特性[7]。CNN 网络可以通过反向传播算法进行训练，比其他人工规则的、单向的神经网络更容易训练，需要优化的参数少得多，这使得卷积神经网络算法在心率检测领域得到广泛研究[8]。

尽管 CNN 对于心律识别任务有诸多优势，但目前将模型部署到计算资源及内存空间有限的可穿戴设备中仍受很多限制。

一方面在于现有的 CNN 模型推理计算需要消耗大量功耗。作为计算量和访存次数极大的模型，CNN 的推理过程包含成千上亿次计算与访存操作，这为硬件资源有限的可穿戴设备带来巨大的功耗负担。

另一方面，保存 CNN 模型需要更大的内存空间。CNN 模型往往通过增加网络深度来提升网络性能，这类设计方法导致权重参数数量和模型大小快速增加[9]。将这些拥有众多参数的模型部署到内存空间有限的可穿戴设备中将是巨大挑战[4]。

因此，低功耗的可便携式智能心率检测设备的研究具有重要的价值，在远程医疗，个人健康监控等领域也有着广泛的应用前景。

1.2 国内外研究现状

1.2.1 生理信号检测算法

对于心电图、脑电图、肌电信号等生理信号，目前主流的检测算法主要分为传统的数据分析算法和神经网络算法两种。生理信号检测算法发展大致经过了三个阶段：

在 20 世纪 90 年代之前，由医生的人工经验演化而来的数据分析算法，被称为经验算

法。这种算法主要依靠医生的临床经验，特征提取规则较为复杂。同时结合了小波变换等数据处理算法[10]，在部分患者身上取得了良好的诊断正确率，但是同样存在泛化能力差，不稳定，算法复杂等局限性，无法满足现代社会对于健康诊断的便捷，通用性强的高要求。由传统算法向人工智能算法发展是生理信号检测算法的必然发展趋势。

到了 90 年代中期，随着人工智能和深度学习等概念的提出和基础机器学习算法的出现，生理信号检测算法领域的研究人员开始聚焦支持向量机(Support Vector Machine, SVM)、随机森林(Random Forest, RF)等人工智能算法，并在部分应用场景西下展现了远优于传统算法的优秀特性[11]。

21 世纪 10 年代之后，人工智能算法百花齐放，诸如 CNN，RNN 等算法在计算机视觉、语音识别等领域展现了得天独厚的优势，生理信号检测算法研究人员也开始关注深度学习、神经网络在心电图等领域的应用前景。更多的类似 CNN、LSTM 等算法开始得到研究应用，同时也出现了多种算法模式混合的检测模型[12]，生理信号检测算法出现了更多的可能性和方法，进一步引领了发展方向。

国外在生理信号检测算法领域的研究开展较早，也较为充分。巴西 Dias Felipe Meneguitti 等人[13]对传统特征提取方式做出了改进。系统采用多特征组合，分别提取了三组特征：RR 间期、信号形态、高阶统计量。提出的分类系统实验结果表明，对 MIT-BIH 数据集中的 N 类、S 类和 V 类的灵敏度分别为 93.7%、89.7%和 87.9%。提出的方法对比传统特征提取方法有明显提升但相比深度学习的方法仍有不足。

波兰克拉科夫工业大学的 Paweł Pławiak 等人[7]设计一种基于深度学习的新方法，以高效快速地对心律失常进行分类。使用了新的用于长持续时间的心电信号片段的 CNN 模型，没有信号滤波、特征波形检测和分割等人工方法，是一种将分类、特征提取和选择阶段相结合的端到端结构。新模型在 MIT-BIH 数据集上表现优异，在 17 分类问题中识别准确率为 86.67%。

吉林大学的研究者[14]在论文中提出了一种非端到端的方法，系统应用主成分分析网络(PCANet)对含噪心电信号进行特征提取。支持向量机(SVM)作为分类器，在 MIT-BIH 心律失常数据库中识别出五种不平衡原始和无噪声的心电图，验证了算法的有效性，其正确率分别达到 97.77%和 97.08%，该方法分类结果显示了较高的准确率，说明该方法是具有一定的噪声鲁棒性和数据适用性。

表 1.1 中列举了国内外具有代表性的生理信号检测算法。代表了当前生理信号检测算法的几个发展方向，目前，国内在基于卷积神经网络的生理信号检测算法领域的研究仍然

缺乏，具有重要的研究价值。

表 1.1 国内外典型的生理信号检测算法

作者	研究单位	算法	说明
Dias Felipe Meneguitti	Universidade de São Paulo	多特征提取	将多特征应用于传统特征提取中。
Paweł Pławiak	Cracow University of Technology	CNN	端到端的方法，没有滤波、特征波形检测等数据处理手段。
Yang W, Si Y, Wang D,	Jilin University	PCANet, SVM	非端到端的方法，两阶段算法，主成分分析网络提取特征，SVM 作为分类器。

1.2.2 神经网络硬件加速器

随着人工智能算法的发展，越来越多的研究人员也开始关注针对人工智能算法的硬件加速器的研究，其中 ASIC 和 FPGA 的硬件实现方式备受推崇。

早在上世纪 90 年代，贝尔实验室就推出了第一款人工智能加速器芯片 ANNA，已经在手写数字集上初步完成了算法的推理能力。但由于当时研究水平和科技能力的限制，这款芯片还是有诸多不足。

随着英伟达开源了 NVDLA 标准化框架，目前的解决方案是使用图形处理单元(GPU)集群作为通用处理器(GPGPU)，但 Venieris, Stylianos I 等人[15]开创性的提出使用现场可编程门阵列(FPGA)，并与上层深度学习软件相结合。FPGA 的可重构性使得其更易于构建和部署模型。由于 FPGA 架构是灵活的，这也可以使研究人员能够探索超出 GPU 等固定架构可能的架构优化。同时，FPGA 功耗低的高性能特点，对便携式生理信号检测系统等的嵌入式应用至关重要。

对于卷积神经网络，Maurice Peemen 等人[16]提出了以内存为中心的设计方法，如图 1-2 所示，可以实现显著的性能。这种增加主要是由于数据访问模式的效率提高所致。通过使用具有灵活的内存层次结构的 HLS 加速器模板，确保支持广泛的 CNN 配置。该加速器架构大大减少了用于嵌入式视觉平台的高效 CNN 加速器的开发时间。

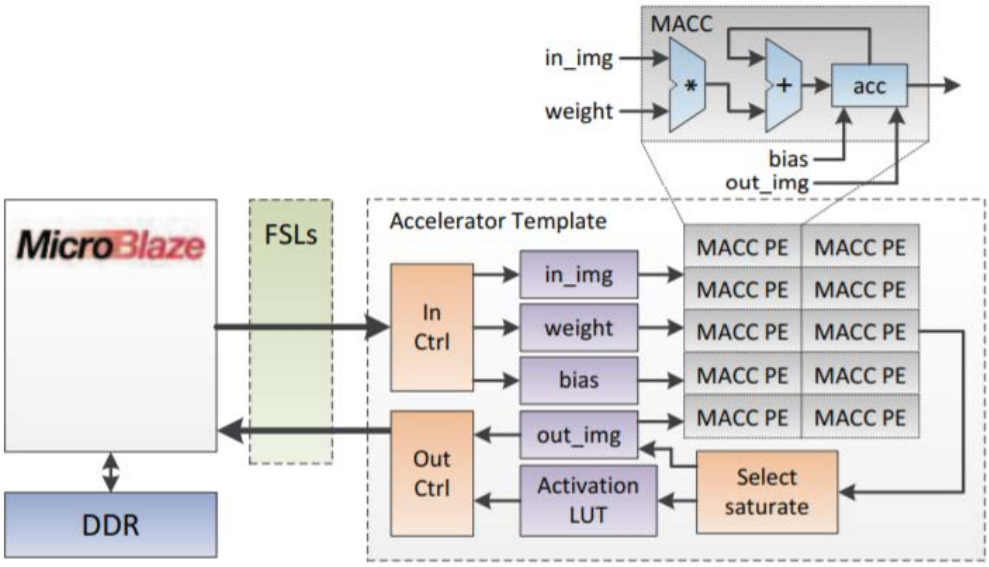


图 1-2 CNN 加速器连接到主机处理器进行控制。

国内学者近几年也开始关注基于神经网络的心电信号检测算法硬件加速器。来自浙江大学的 JIAQUAN WU 等人[17]提出了一种用于多分类心电信号检测的加速器，如图 1-3 所示。该框架可以同时以 BLSTM 和 CNN 网络模型为推理对象，同时分类算法以较小的网络规模实现了高精度的心律失常检测，并设计了版图。基于重用的硬件体系结构大大加快了推理过程，降低了功耗，这对于可以长期携带的智能心率检测装置非常有效。

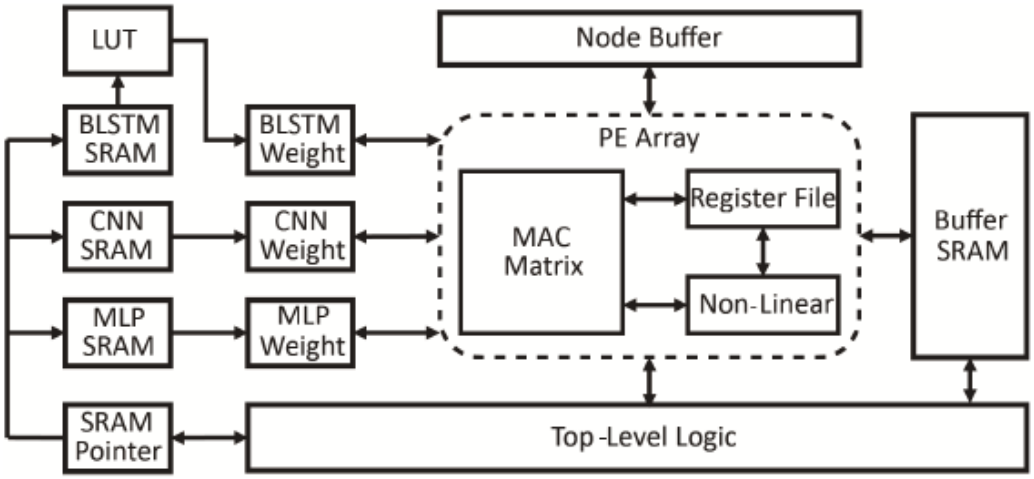


图 1-3 处理器总体架构，BLSTM、CNN 和 MLP 层的权重存储在静态随机存储器中，并独立地读出到相应的权重缓冲器。

表 1.2 中列举了国内外具有代表性的神经网络硬件加速器。目前，国内对于基于卷积神经网络的生理信号检测算法硬件加速器仍然缺乏，虽然在图像处理等领域，研究学者提出了众多卷积神经网络加速器架构，但是对于可以应用于生理信号检测的硬件加速器电路

研究仍缺乏，具有重要的研究价值。

表 1.2 国内外典型的神经网络硬件加速器

作者	研究单位	实现方式	说明
Venieris, Stylianos I	Imperial College London Imperial Coll London	FPGA	集成 FPGA 到现有的深度学习环境中，以提供性能、功耗和可编程性之间的平衡。
Maurice Peemen	Thermo Fisher Scientific	ASIC	针对卷积神经网络，大大提高了数据访存效率。
JIAQUAN WU	Zhejiang University	ASIC	针对 BLSTM 和 CNN 两种模型的可重用架构。

1.3 研究内容及设计指标

1.3.1 研究内容

本课题针对心律失常问题，提出了一种基于卷积神经网络的新型生理信号检测算法，并结合 CNN 网络模型的特点，完成了对应的硬件加速器电路设计，使其在保证算法准确率的前提下，减少功耗和硬件资源的使用。

其主要研究内容包括：

(1) 提出了一种基于卷积神经网络的新型生理信号检测算法，在保证算法准确率的前提下，优化了网络模型，使其在硬件上更好的实现。

(2) 针对提出的网络模型进行了参数压缩。将原先的浮点数模型进行量化，大大减小了权重参数所占内存空间。

(3) 针对提出的新型生理信号检测算法，设计了对应的硬件加速器电路。优化了加速器在访问内存和计算方面的功耗。

1.3.2 设计指标

本课题的主要目标是寻找一种新型生理信号检测算法，在保证算法准确率的前提下，实现功耗和网络性能的平衡，完成算法到硬件加速器电路的映射。主要的功能和性能指标包括：

(1) 提出一种基于卷积神经网络的新型生理信号检测算法。

(2) 设计算法对应的低功耗卷积神经网络硬件加速器电路。

(3) 优化模型，使其模型大小压缩率达到 50%以上。

(4) 设计的生理信号检测卷积神经网络算法在 MIT-BIH 数据集上的识别准确率不低于 90%，硬件加速器电路的识别准确率不低于 85%。

(5) 完成加速器电路设计，加速器电路部分功耗不高于 1W。

1.4 论文组织结构

本课题主要针对心律失常问题提出了一种新型基于卷积神经网络的生理信号检测算法，并搭建了对应的加速器电路模型，完成了从算法到电路的映射。最终完成功能仿真和性能评估。本文将分为 6 个章节对工作内容进行叙述。

第一章为绪论，针对生理信号检测算法及其硬件加速器的研究背景、意义、国内外研究现状和本文研究内容。

第二章为生理信号检测方法与卷积神经网络概述。首先，分析了传统生理信号检测方法和基于人工智能的检测方法。最后，解释了卷积神经网络的相关原理。

第三章为生理信号检测算法设计及参数压缩。首先，介绍了本文所用到的数据集。接着，提出了一种新型生理信号检测算法以及算法测试结果。最后，介绍模型压缩方法和量化原理，并对量化模型进行了分析。

第四章为生理信号检测卷积神经网络硬件加速器电路设计。针对提出的新型生理信号检测算法，完成了卷积神经网络加速器电路的设计工作，并对各个子模块的设计方法进行了详细描述。

第五章为仿真与测试。对加速器电路完成了功能仿真测试，并以完成了电路性能评估以及最终的实验结果。

第六章为总结与展望，对全文进行系统性的总结，并提出可行的改进方向。

第二章 生理信号检测算法与卷积神经网络概述

本章主要对生理信号检测算法与卷积神经网络相关技术进行讨论，首先分析了传统生理信号检测算法的实现步骤；然后分析了基于神经网络的生理信号检测算法，以 CNN、BPNN、RNN 为实例，详细分析了基于神经网络的生理信号检测算法相对于传统算法的优势；最后对卷积神经网络原理进行论述。

2.1 传统生理信号检测算法

传统的生理信号检测算法通常分为三个步骤：数据预处理，特征提取和模式分类[18]。

以最常见的心电信号为例。心电信号的采集通常是通过专业的心电记录仪获取的。心电记录仪在采集心电信号的同时往往会伴随一些噪声信号。ECG 信号中主要的噪声信号有心肌噪声、信道噪声等。针对这些噪声，目前主流去除噪音的方法有：滤波法、共模抑制法等。滤波法主要采用低通滤波器和自适应滤波器，共模抑制法的原理是很多噪声是共模干扰所产生的，可以利用差分放大器的结构对称性来消除这种共模噪声。

信号预处理完成之后，就是提取心电信号的特征。当人体心肌细胞细胞膜内外离子的浓度不同，处于极化状态时，一旦收到来自起搏细胞的激动，这种极化状态会暂时消除，在心电图上称之为“去极化”。在“去极化”过程中，钠离子内流，钾离子外流，形成反向电势差，电位突变产生脉冲信号。P 波代表心房的去极化。QRS 复合波代表心室的去极化。T 波代表心室的复极化，采集到各心肌细胞的动作电位叠加后形成如下图 2-1 所示的心电信号。

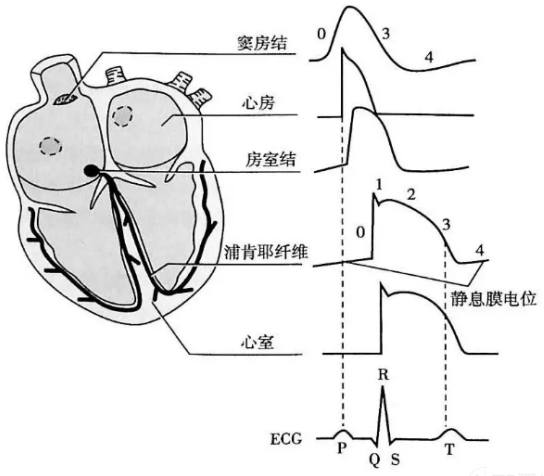


图 2-1 体体表采集到各心肌细胞的动作电位和叠加的 ECG 信号

在提取波形特征时，各个波形的变化情况最受到重视[19-21]，例如，当 QRS 波变大变宽时，可能发生了室性早搏；ST 段抬高时，可能发生了心肌梗死。这样，通过最直观的波形变化，结合医生的经验，可以进行疾病的诊断。常用的形态特征有：①P 波振幅 ②QRS 波振幅 ③T 波振幅 ④PR 间期 ⑤QRS 间期 ⑥QT 间期 ⑦ST 间期 ⑧PR 段水平 ⑨ST 段水平 ⑩RR 间期，如图 2-2 所示。这些形态特征通常可以通过 Hermite 变换，小波变换[10]或离散余弦变换[22]等算法得到。

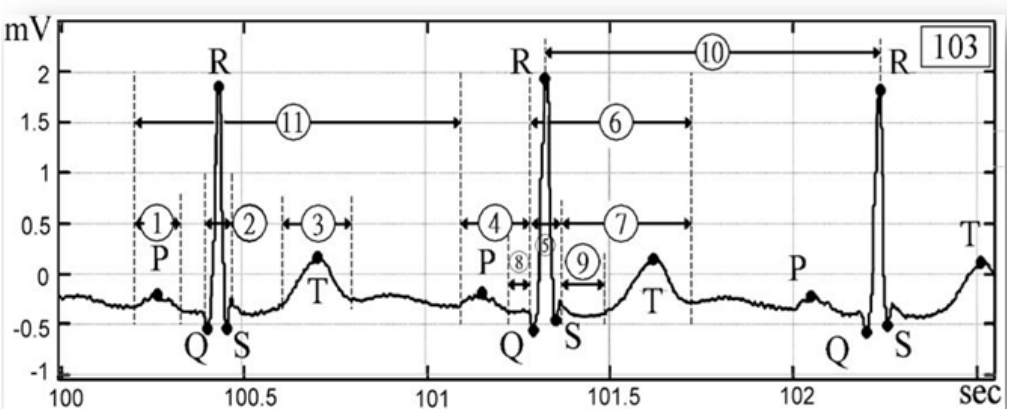


图 2-2 心电图形态特征

提取心电图特性后，便是根据特定的检测和识别任务设计分类器。在现有的研究中，研究人员主要使用经验公式分类和机器学习分类算法。尽管依靠经验公式分类，规则复杂，但是在个体患者上取得了良好的准确率，同时具有医学上的可解释性。机器学习的分类算法如线性和支持向量机(SVM)、随机森林(RF)、贝叶斯网络等。文献[23]在对 3 类 ECG 信号的小波变换特征和 RR 波间隔上采用线性判别分析算法进行分类检测，取得了 93%的平均正确率。Nasiri 等人[24]提取的 22 个形态特征，使用 SVM 算法对其分类，并在 4 类心律失常检测中达到了 93%的正确率。

2.2 基于神经网络的检测算法

过去五年中，生理信号检测算法的重大进步在很大程度上是由神经网络所带来的。神经网络是由多个处理层组成的计算模型，每层都可以学习上一层输出数据越来越抽象的表示，最终起到特征提取的作用。目前，神经网络算法已经在语音识别[25]、图像识别[26]和医疗应用[27, 28]方面展现了最出色的效果。神经网络算法的识别模式和从原始数据中学习关键分类特征的能力是传统算法不能做到的，无需利用手工规则对数据进行预处理，这种特性

使其特别适合解释 ECG 数据。此外，由于神经网络算法性能往往随着训练数据量的增加而增加，因此这种方法会随着 ECG 数据的广泛数字化而进一步提高诊断准确率。

此外，来自山东大学和华中科技大学的学者研究了过去十年(2010 年至 2019 年)中常用的神经网络的比例关系[29]，使用的各种网络的比例如图 2-3 所示。可以看出，在这些网络中，CNN 是最受欢迎的。其他三个网络，PNN、RNN 和 BPNN 的使用比例大致相同。由于心电信号也像图像信号一样具有局部相关性和平移不变性的特点，在图像领域表现优异的 CNN 也在心电信号检测上成功地应用。

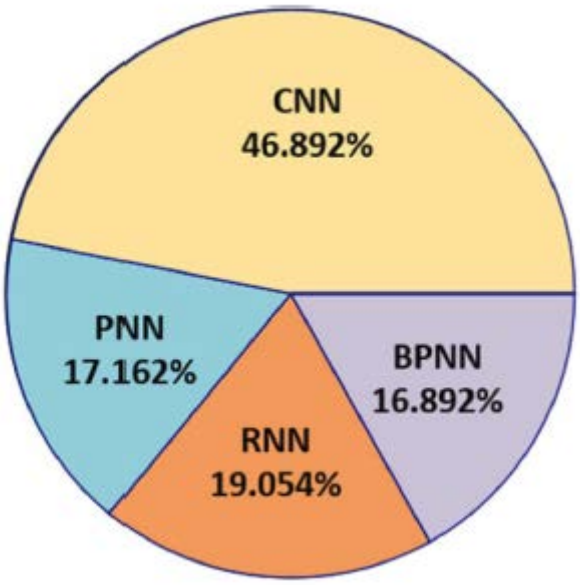


图 2-3 近十年(2010-2019 年)几种常用的心电处理 AI 模型的比例关系。

来自新加坡的研究者[30]提出了一种由CNN模型组成的端到端的冠心病自动诊断系统。算法模型结构包括四个卷积层，四个最大池层和三个全连接层，该自动诊断系统可以通过一段 5 秒持续时间的心电信号来诊断冠心病。实验结果显示，自动诊断系统的准确率为94.95%，灵敏度为 93.72%。临床医生可以通过该系统对异常心电信号做出更准确可靠的冠心病诊断。

反向传播神经网络也在心电信号诊断中有着应用价值[31]。巴基斯坦的学者使用反向传播神经网络(BPNN)作为分类器，实现了对健康人和拥有心脏病潜在风险患者进行分类。使用的心电图数据来自 Physio-bank 提供的 PTB 数据库。该算法通过提取 ECG 信号中的时域特征，如 T 波振幅、ST 间期等，接着对获得的时域形态特征进行主成分分析(PCA)，最终输入 BPNN 网络进行分类。在检测中，发现 BPNN 对心电信号分类的灵敏度和特异性

分别为 97.5%和 99.1%。

RNN 模型通常用于语音识别等一维信号，同样 RNN 在长时间的心电信号分类中也表现优异。土耳其学者提出了一种方法[32]，用于心电图的自动诊断，并在测试中表现优异。方法将递归神经网络(RNN)用作检测心电图信号的基础，对 Physio-bank 数据库获得的四种心电图节律(正常跳动、充血性心力衰竭中风、心室心律失常跳动、心房颤动中风)进行分类。算法分为两个阶段进行：将 ECG 信号形态特征输入 RNN 网络，使用 Levenberg-Marquardt 算法训练 RNN。实验结果显示，RNN 在这四个分类上，实现了高分类精度。

随着神经网络算法的快速发展，在心电信号诊断领域已经开始逐步取代传统生理信号检测算法，开始成为主流研究方向。基于神经网络的生理信号检测算法的技术研究方案和医疗诊断系统也不断被推出，其利用先进的神经网络算法，对便携式健康监测设备采集的心电信号进行分析处理，提供了全面，精准和科学的医学诊断结果。

2.3 卷积神经网络原理

卷积神经网络(Convolutional Neural Network, CNN)是一种常用的神经网络模型，也是当前在图像识别领域最常用的算法之一。卷积神经网络中不同的操作通常被称为“层”。如图 2-4 所示，CNN 模型通常由卷积层、激活函数、池化层和全连接层组成。

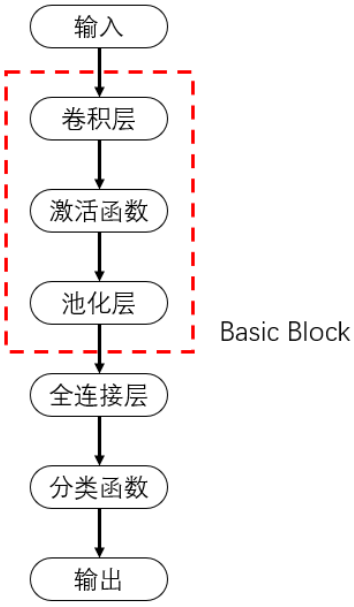


图 2-4 卷积神经网络基础结构

2.3.1 卷积层

卷积层是卷积神经最核心的功能层。顾名思义，卷积层的功能就是进行卷积计算。卷积层有两种主要参数：卷积核 W 和偏置 b。这里以一组一维数据进行卷积操作说明。如图 2-5 所示，输入数据 x 为(1, 2, 3, 1, 2, 3)，卷积层卷积核参数 W 为(0, 1, 2)，第一次卷积计算结果为$0 \times 1 + 1 \times 2 + 2 \times 3 = 8$，后三组计算同理。最终可以得到输出 y 为(8, 5, 5, 8)。

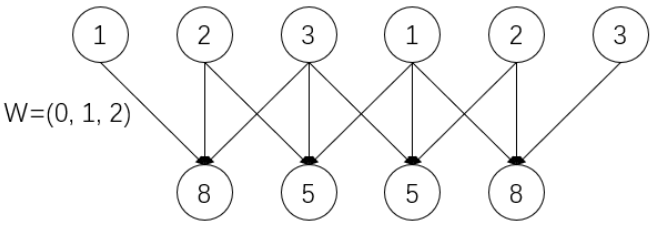


图 2-5 一维卷积操作

数据处理中经常用到卷积操作，卷积运算作为一种线性相关运算，实质上是对输入数据的一种特征提取手段。一维卷积运算的具体公式如下：

$$y_m = \sum_{i=0}^{I-1} x_{m+i} W_i + b_m \,, \qquad m \in [0, M) \tag{2.1}$$

公式(2.1)中 y 为该卷积层的卷积输出；x 为大小为 M 的一维输入数据；W 是大小为 I 的卷积核参数，通常称之为权重；b 为偏置。

2.3.2 激活函数

激活函数通常存在于层与层之间。在卷积神经网络中，一层的激活函数定义了该层在给定的输入下的输出。通常输入达到一定大小时，函数所在层就存在输出，称之为这层或该神经元被“激活了”，将该函数称之为“激活函数”。激活函数包括 ReLU, Sigmoid, Tanh 等。

ReLU 函数是深度学习中最流行的一种激活函数，相比于 sigmoid 函数和 tanh 函数，ReLU 只有当输入为正的时候才有输出，不存在梯度饱和问题，并且只存在线性关系，它的计算速度快得多。ReLU 激活函数公式如下，图像如图 2-6 所示。

$$y = \max(0, x) \tag{2.2}$$

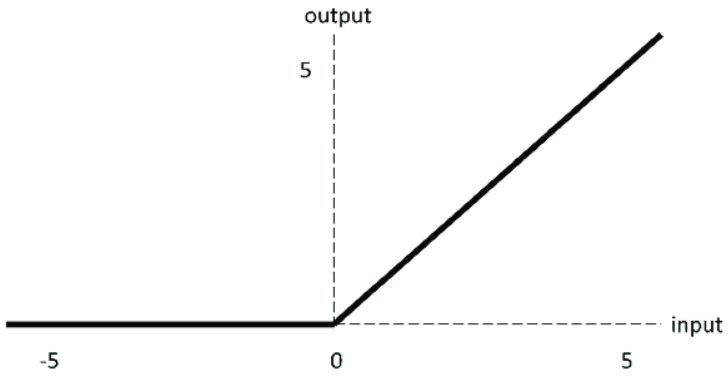


图 2-6 ReLU 激活函数

Sigmoid 函数通常用于二分类问题中。Sigmoid 函数的输出在 0 到 1 之间，相当于对每一个输入数据进行了归一化。在二分类问题中可以进行非常明确的预测，即非常接近 1 或 0。Sigmoid 激活函数公式如下，图像如图 2-7 所示。

$$y = \frac{1}{1 + e^{-x}} \tag{2.3}$$

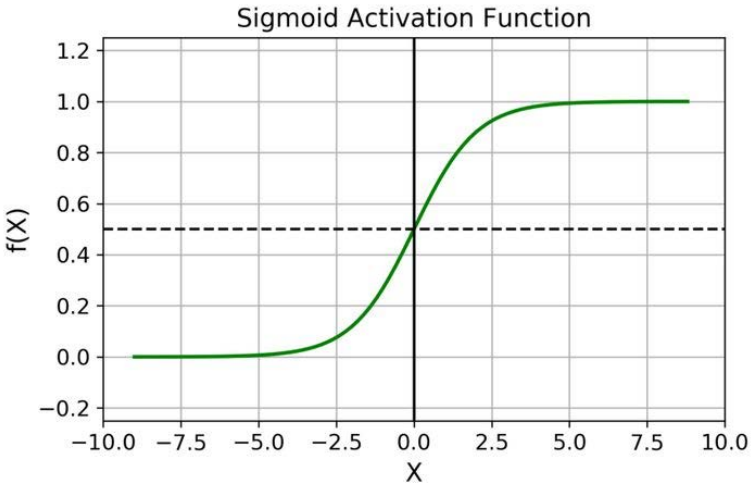


图 2-7 Sigmoid 激活函数

双曲正切 Tanh 激活函数与 Sigmiod 激活函数类似也是 S 型。但是其输出范围却是-1 到 1，当输入较小或者较大时，激活函数 Tanh 曲线较为平缓，这一点不利于权重的更新。与 Sigmoid 激活函数不同的是，当输入为负的时候，其输出也为负，当输入为 0 时，输出也为 0，并且通常用于二分类问题中的隐藏层，而非输出层。值得注意的是，Tanh 与 Sigmiod 激活函数都包含了指数函数，当映射为电路时，实现过程较为复杂，并且存在近似计算问题，这都将影响最终电路的结果。Tanh 激活函数公式如下，图像如图 2-8 所示。

$$y = \frac{2}{1 + e^{-2x}} - 1 \tag{2.4}$$

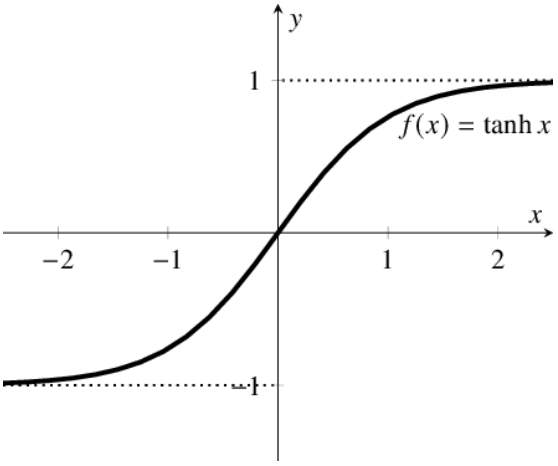


图 2-8 Tanh 激活函数

2.3.3 池化层

对于图像或者心电图这种具有连续性的输入数据而言，相邻的输入数据通常具有相似的值，因此卷积层的输出通常具有类似的值。换而言之，卷积层输出中的大部分信息都是重复的。池化层解决了这个问题，池化层所做的是减少输出值的数量。池化层通常分为最大池化层，平均池化层或最小池化层，即执行选取最大的值，计算平均值，选取最小值的操作。图 2-9 展示了一组一维数据经过池大小为 3 的最大池化层的效果。

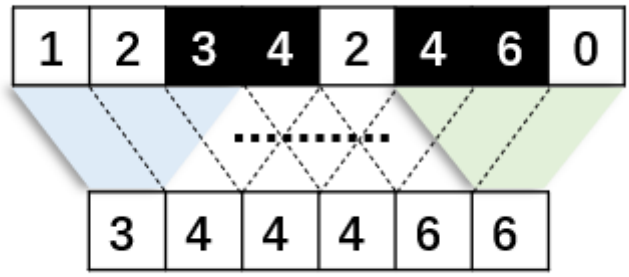


图 2-9 最大池化层原理

除此以外还存在一种特殊的池化层，全局平均池化层。在一个卷积神经网络模型中，最后输出之前通常存在一个全连接层和 softmax 激活函数。全连接层的参数极多，这使得模型变得非常臃肿。新加坡学者 Lin, Min2014 年提出了使用全局平局池化层(global average poolilng, GAP)代替全连接层的方法[33]。全连接网络的目的是减少特征图维度，进而输入到 softmax 激活函数中进行分类，用池化层来代替全连接层也能起到同样的降维效果，及一个通道的特征图平均为一个数。GAP 的结构如图 2-10 所示。

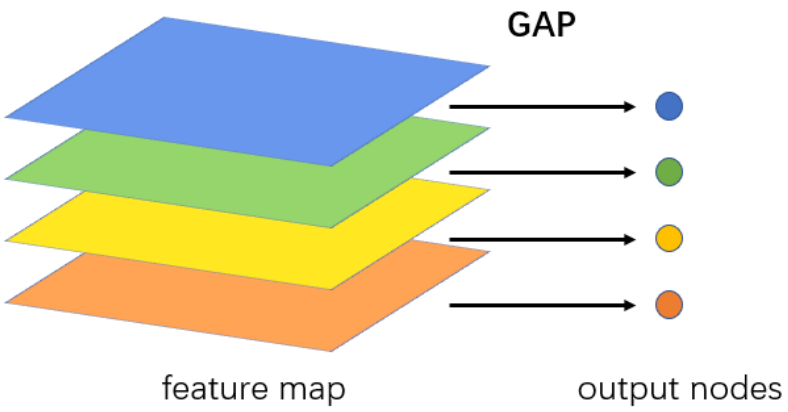


图 2-10 全局平均化池原理

2.3.4 全连接层

全连接层的所有节点都与上一层的每一个节点相连，它可以把一个特征图的所有特征综合起来映射到样本标记，起到数据降维的作用。如图 2-11 所示，假设上一层的特征向量有 m 维，经过全连接层之后得到一个 n 维的列向量输出，对应 n 个分类目标，那么全连接层的参数数量为$m \times n$个。可以看出全连接层的参数数量远多于卷积层，计算量比较大。

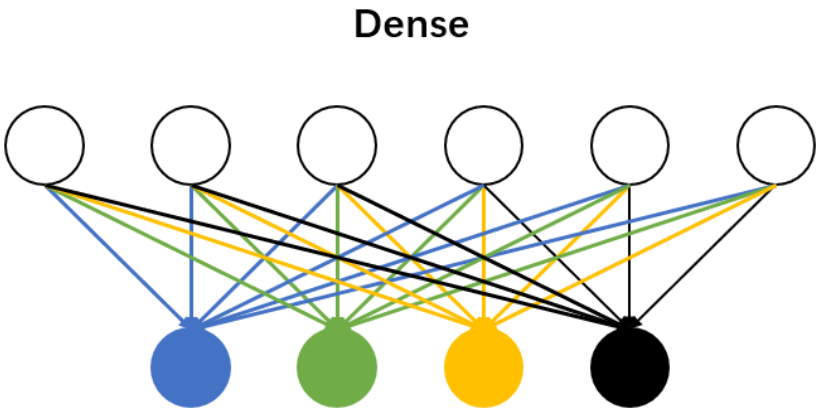


图 2-11 全连接层原理

与全连接层经常配合使用的激活函数为 softmax。Softmax 常用于多分类问题中，可以将全连接层的输出映射到(0, 1)的区间里，最终的输出可以解释为每种分类的概率。Softmax 激活函数公式如下。

$$S_i = \frac{e^j}{\sum_j e^j} \tag{2.5}$$

公式(2.5)中，i 表示输入向量共 i 维，即目标分为 i 类，最终 i 个输出概率总和为 1。最大的S_i就对应概率最大的节点，也就是最终的预测结果。

2.4 本章小结

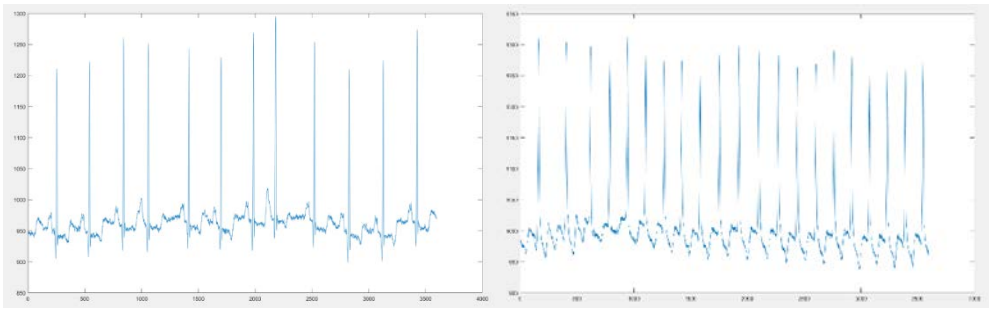
本章首先从传统的生理信号检测算法入手，详细分析了以特征提取和数据分析为核心的传统生理信号检测算法的结构特点；然后分析了基于神经网络的生理信号检测算法，以 CNN、BPNN、RNN 等网络模型为典型代表，并与传统算法进行对比分析，突出了神经网络算法的优势；最后，详细论述了卷积神经网络的理论原理，包括卷积层、激活函数、池化层和全连接层的实现方式，为下一章节的基于卷积神经网络的新型生理信号检测算法的提出打下理论基础。

第三章 生理信号检测算法设计及参数压缩

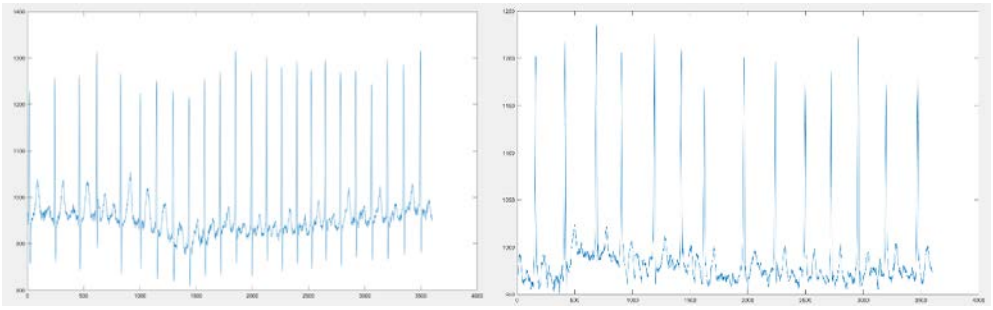
本章节主要由针对心电信号的基于卷积神经网络的新型生理信号检测算法及算法模型参数压缩两部分组成。针对 MIT-BIH 心律失常数据集，本课题提出了一种直连型卷积神经网络结构，其中包含了 6 层基础卷积层和 2 层全连接层，并针对该网络模型进行了优化；同时采用量化的方式，对网络模型参数进行了压缩，并分析了最终的实验结果。

3.1 数据集

本文使用 Pławiak 的数据集[7]进行网络训练和性能评估。数据集中的 ECG 信号来自 PhysioNet 的 MIT-BIH 心律失常数据集，共包含 48 条长度为 30 分钟的双导联 ECG 数据，所有的 QRS 波都有对应的标注，所有的 ECG 数据经 0.1-100Hz 的带通滤波器滤波后，在 360Hz 下进行采样。如图 3-1 所示，列举了四种来自 modified limb lead II 具有代表性的心率类型。



(a)　　(b)



(c)　　(d)

图 3-1（a）Atrial premature beat 心率信号（b）Supraventricular tachyarrhythmia 心率信号

（c）Atrial flutter 心率信号　　（d） Atrial fibrillation 心率信号

该数据集的特点为：(1)共有 1000 条数据，每条数据的长度为 10 秒，不同样本之间互不重叠；(2)信号共来自 45 名患者：包括 19 位女性（23-89 岁）和 26 位男性（32-89 岁）；(3)信号包含 17 种类别：正常窦性心律，起搏器节律和 15 种类型的心律失常，每种类型至少包含 10 段信号；(4)信号均来自 modified limb lead II。心律类型分布如下表 3.1 所示：

表 3.1 心律类型分布

No.	Class	Instances Number
1	Normal sinus rhythm	283
2	Ventricular tachycardia	10
3	Idioventricular rhythm	10
4	Ventricular flutter	10
5	Fusion of ventricular and normal beat	11
6	Left bundle branch block beat	103
7	Right bundle branch block beat	62
8	Second-degree heart block	10
9	Pacemaker rhythm	45
10	Atrial premature beat	66
11	Atrial flutter	20
12	Atrial fibrillation	135
13	Supraventricular tachyarrhythmia	13
14	Pre-excitation (WPW)	21
15	Premature ventricular contraction	133
16	Ventricular bigeminy	55
17	Ventricular trigeminy	13
	Total	1000

在实验阶段，选取了 70%的数据作为训练集训练网络，剩余的 30%的数据作为测试集来对训练后的网络性能进行测试，如图 3-2 所示。考虑到数据集中不同类型的样本数量差异巨大可能会导致训练效果不佳，于是对训练集的数据进行了样本均衡，对样本数较少的种类进行过采样，使不同类型样本数量趋近一致。

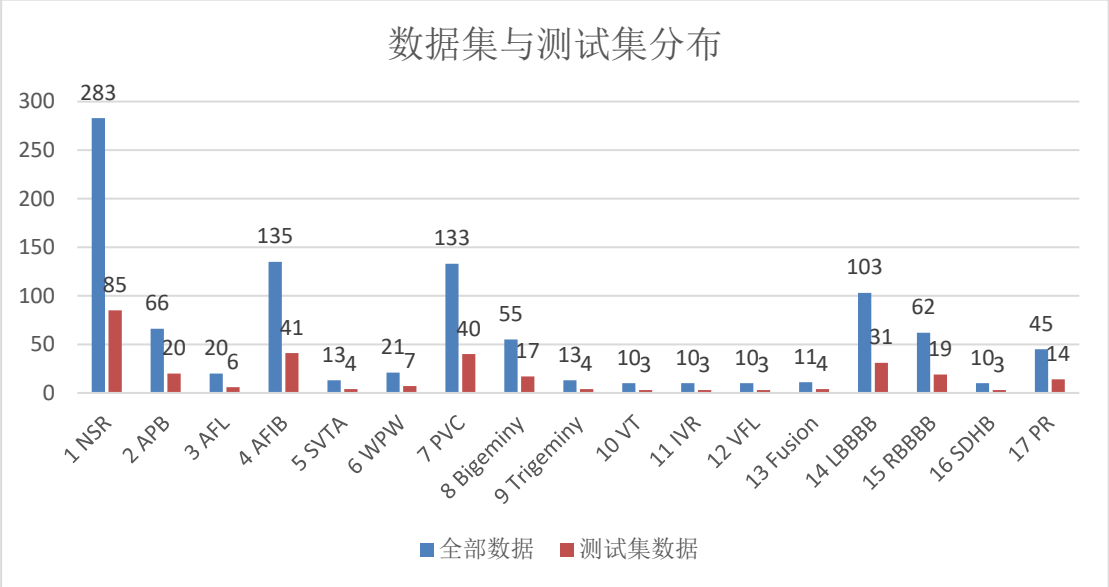


图 3-2 数据集与测试集数量分布情况

3.2 检测算法模型搭建及结果分析

3.2.1 卷积神经网络的搭建

本课题提出了一种针对 ECG 信号的基于卷积神经网络的新型生理信号检测算法，网络训练流程如图 3-3 所示。

首先对原始数据集数据进行过采样处理，对所搭建的 CNN 网络进行参数训练。然后，对网络层参数量化，直到参数的压缩比例和心率检测精度均满足要求。

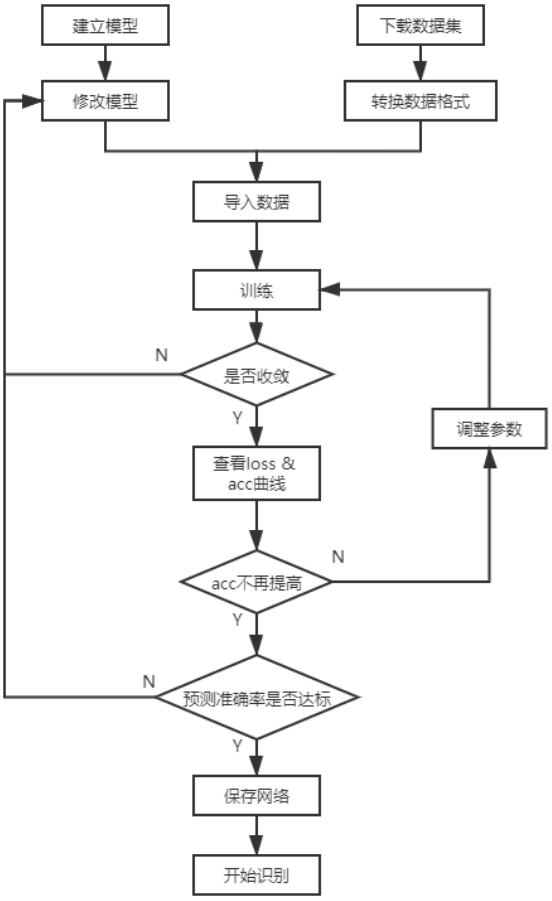


图 3-3 基于 ECG 序列的卷积神经网络模型训练流程

该网络使用一段 10 秒的 ECG 序列（包含 3600 个样本点）作为输入。该网络可以根据输入的信号端到端的输出信号所属的心律类型。这种设计避免了在信号输入网络之前的心拍检测的阶段。尽管许多基于心拍的心率检测网络显示出优越的性能，但由于神经网络的输入通常是固定长度的序列，而这类网络的输入是一段完整的心拍信号且不同心律的持续时间差异很大，因此不仅需要在输入网络之前对输入信号执行 QRS 波形检测，还需要根据检测结果对数据处理。该网络避免了对心拍的检测，能够在不丢失相邻心拍特征的同时避免了心拍分割阶段的功耗，进一步简化了系统结构。

本课题所设计的网络结构采用直连型网络结构。该网络由下图 3-4 所示的基本块和两个全连接层串联组成，每个基本块受两个超参数限制：卷积核 k 的大小和卷积核数 m。由卷积层组成的基本块用于提取 ECG 信号特征，全连接层用于综合全局特征进行分类，使用 softmax 作为激活函数包含 17 个输出节点，分别对应属于每种类型心律的概率。考虑到心电信号的形态特征是判断心律不齐的重要特征，研究中还使用大核卷积核对一段较长的临近采样点进行特征提取，以此获得较长时间跨度上心电图信号所包含的信息。

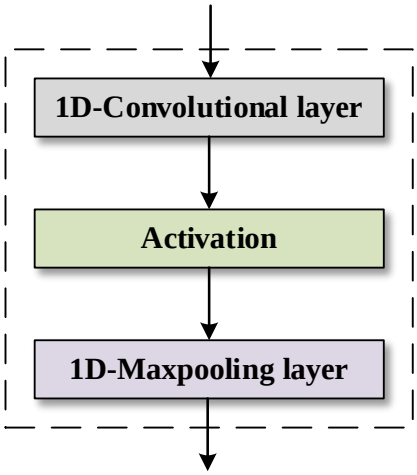


图 3-4 算法模型基本块

在直连型网络模型中，网络的深度可以有效的提升网络的性能，但是网络结构太深会带来更多的计算功耗和内存。先对不同的深度的网络结构与网络识别精度和所需内存的关系进行实验，可以根据网络的识别准确率和内存确定最终的网络拓扑结构。图 3-5 为 8 种由基本块组成的不同深度的直连型网络拓扑结构图。所有配置的结构都由多个基本块和全连接层串联组成，基本块用于提取 ECG 信号特征，全连接层用于综合全局特征进行分类。

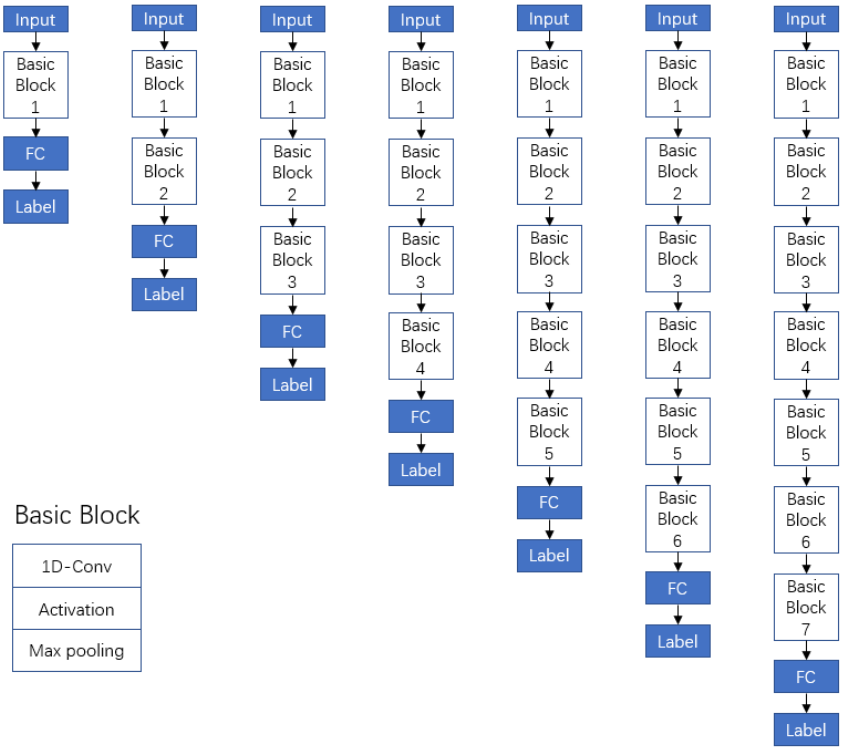


图 3-5 网络拓扑结构图

在使用相同的训练样本对所有结构的神经网络进行 3000 次训练后，选取训练过程中整体准确率最高的模型，即可得到最终需要的网络拓扑结构。

在对不同拓扑结构的网络的整体准确率、网络内存大小进行统计之后，数据如图 3-6 所示。网络识别准确率随着基本块的增加而增大，当基本块的数量为 6 时，识别准确率达到最高，为 94.7%，之后增加基础块数量准确率不再升高。与识别精度曲线不同，由于受到卷积层参数和全连接层参数的影响，内存大小总体呈现先上升后下降的趋势，当基本块为 2 和 7 时，所需内存达到局部最小值，分别为 343KB 和 315KB。

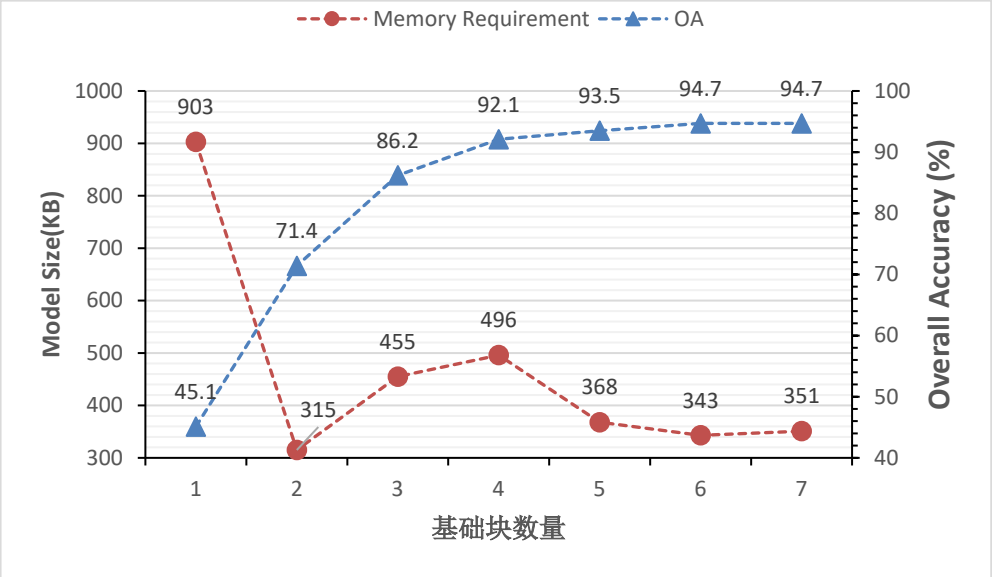


图 3-6 网络精度与网络大小关系图

综上所述，当网络结构由 6 个基本块和 2 全连接层组成时，识别准确率为 94.7%，同时，网络所需内存大小最低为 343KB，是最佳拓扑结构。因此，选择该配置的结构作为本设计最终的模型，并使用 GAP 层来代替第一个全连接层。最终网络结构如图 3-7 所示，网络参数如表 3.2 所示。

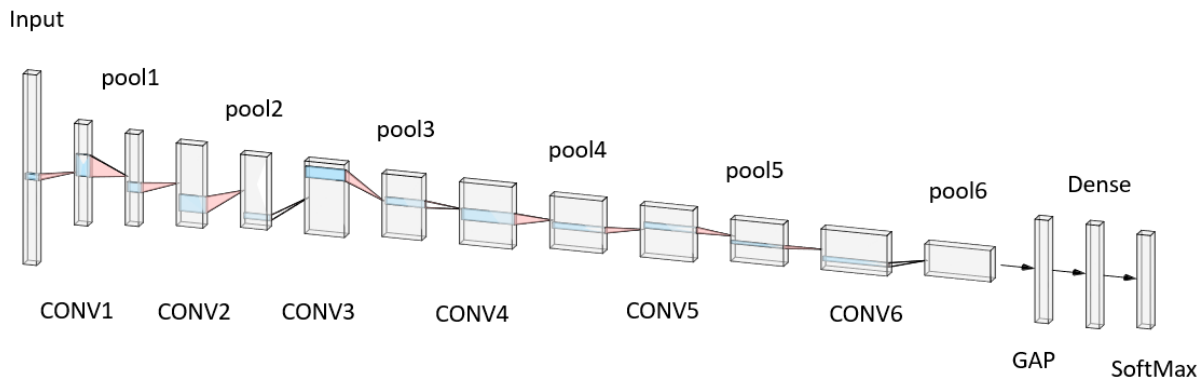


图 3-7 生理信号检测算法神经网络模型图

表 3.2 生理信号检测算法神经网络参数表

层名	CONV1	CONV2	CONV3	CONV4	CONV5	CONV6	GAP	Dense
输入特征图尺寸(IX)	3600	894	220	105	48	22	10	1
输入特征图通道数(M)	1	8	16	32	64	64	72	72
输出特征图尺寸(OX)	1793	442	213	98	45	20	1	1
输出特征图通道数(N)	8	16	32	64	64	72	72	17
卷积核尺寸(K)	16	12	8	8	4	3	-	-
卷积核步长(S)	2	2	1	1	1	1	-	-
权重数量(W)	128	1536	4096	16384	16384	13824	-	1241
乘加操作数量(MAC)	445K	1,301K	1,636K	3,011K	1,290K	461K	-	-

3.2.2 实验结果

本课题将提出的新型生理信号检测算法卷积神经网络模型在 MIT-BIH 数据集上进行训练，训练过程如图 3-7 所示。图中可以看出，经过 500 次训练，新型生理信号检测算法的是被准确率达到了 94.4%。损失函数变化如图 3-8 所示，此时损失函数接近于零，已经收敛。其中橘色为模型在训练集上的表现，蓝色曲线为模型在测试集上的表现，上述曲线均经过平滑处理。

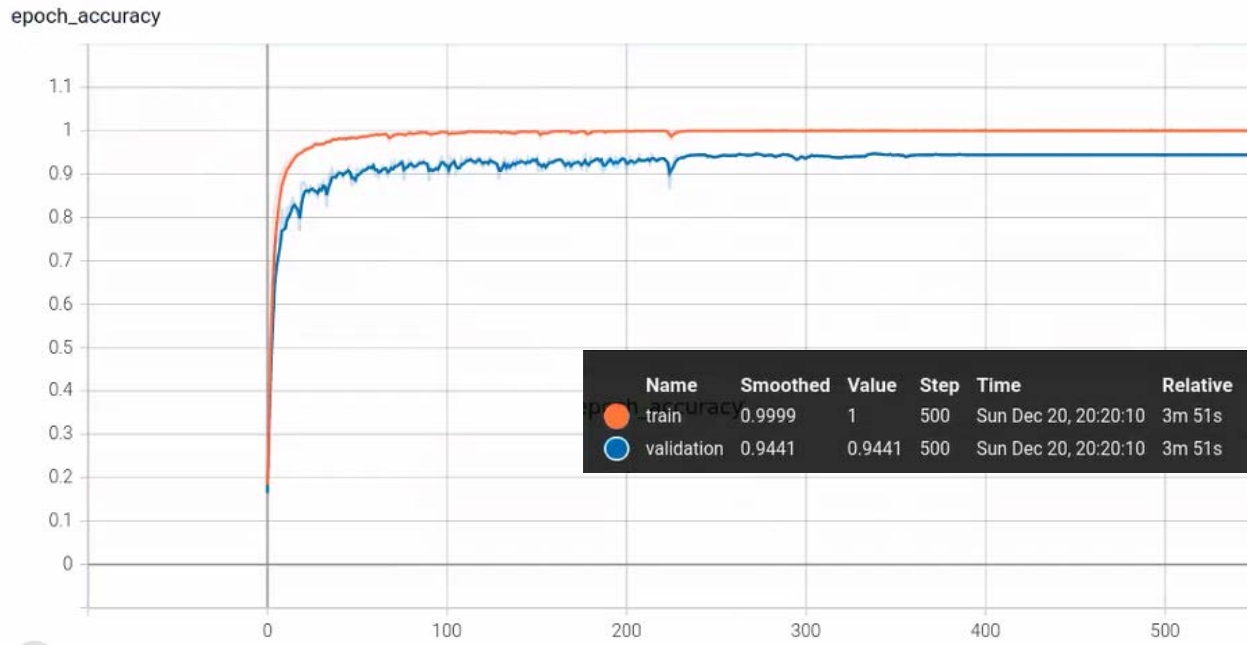


图 3-7 网络训练准确率曲线图

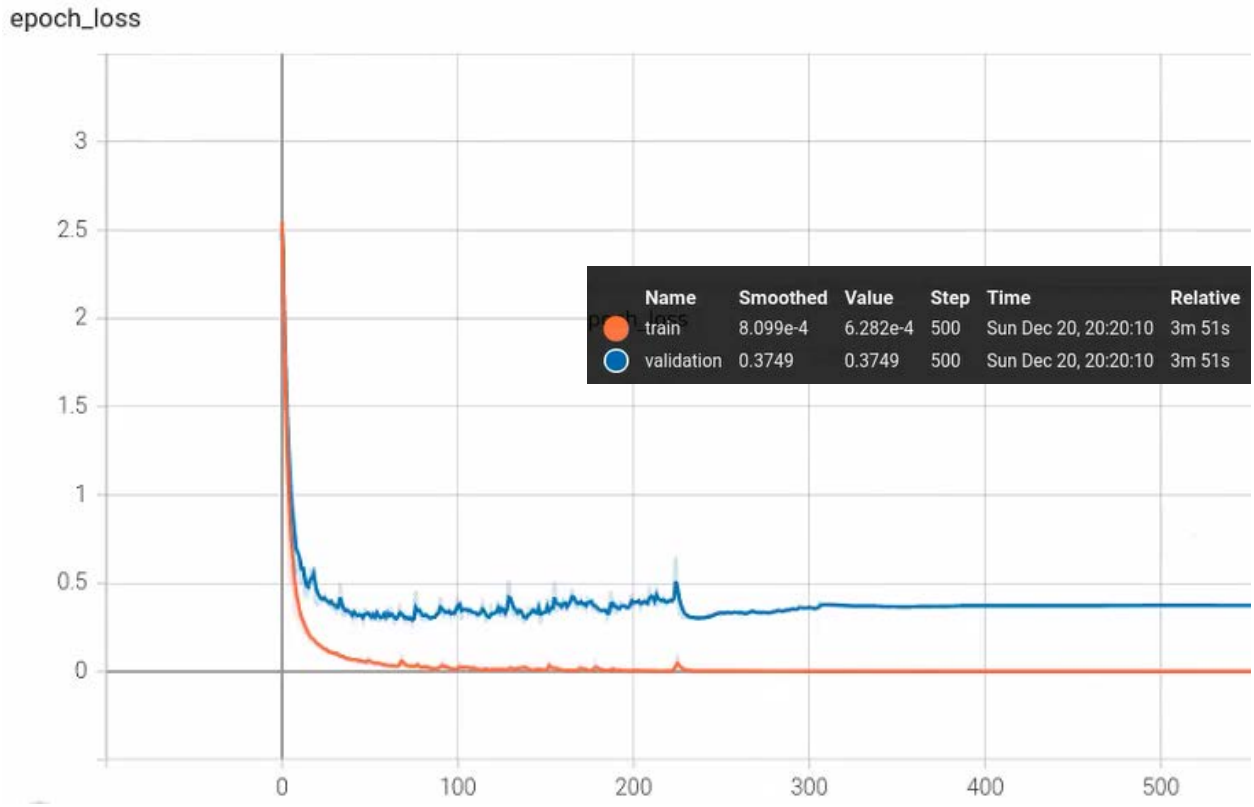


图 3-8 网络训练损失函数曲线图

完成训练之后，以测试集识别准确率为目标，选取最佳模型。如图 3-9 所示，最佳模型出现在第 335 轮训练时。模型训练集准确率达到 100%，测试集准确率达到 94.7%。

图 3-9 最佳网络模型

最佳模型在对 304 个测试集样本进行识别中，该网络正确分类了 288 个样本，实现了 94.7%的测试准确度。同时，对于所有 17 种心律类型的识别，各个类型的正确的分类均占主导；其中，有 8 种心律的识别准确率为 100%，没有误诊。此外，观察混淆矩阵还可发现，通过在训练阶段对训练数据中的低数量类型心律进行过采样操作，该网络对样本总数小于 50 的 10 类少数量样本的识别

准确率达到 96%。图 3-10 为上述神经网络在测试数据中的混淆矩阵。

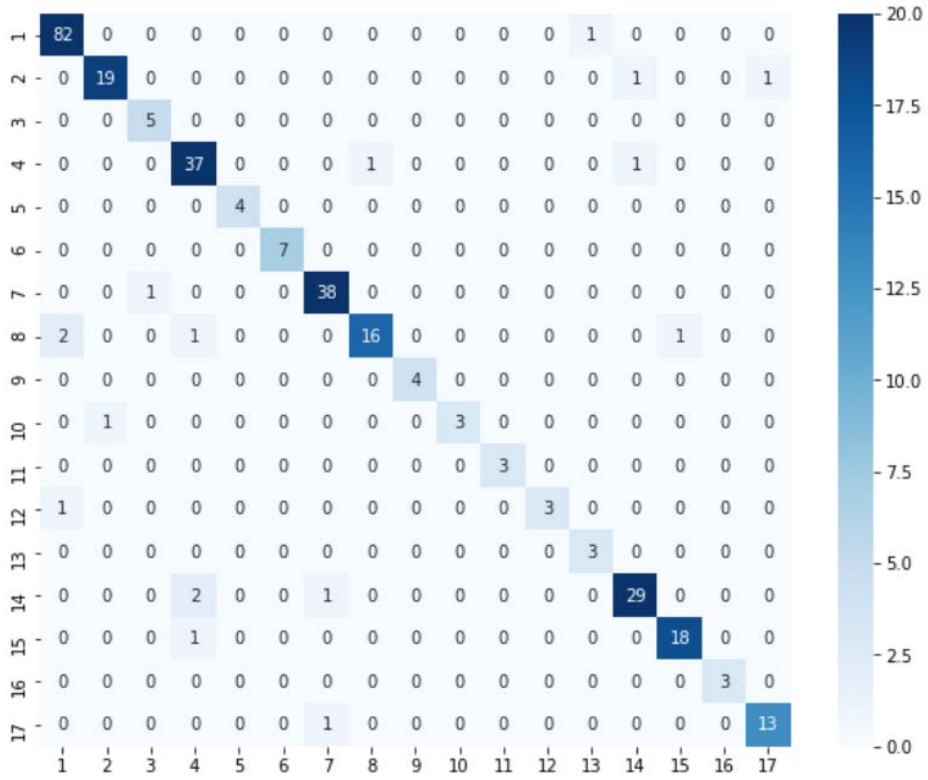


图 3-10 混淆矩阵

表 3.3 对比了在 17 分类问题上表现最好的 Yıldırım 所研究的长期心律失常分类器算法及本文的神经网络在心律失常分类准确率和内存需要，两种算法均在 MIT-BIH 数据集 17 分类上进行了测试。观察表 3.3 可以看出，本文提出的神经网络在心律识别准确率达到了 94.7%。相比 Yıldırım 的算法，整体精度提高了 3.4%。 此外，本文提出的网络结构大小仅为其的 38 分之一。

表 3.3 本文提出的网络与其他文献工作的性能对比

	Sen(%)	Spe(%)	OA(%)	Memory(KB)
Yıldırım[7]	83.91	99.41	91.33	7852
Proposed	93.71	99.52	94.74	209.2

3.3 参数压缩

对于心律失常自动诊断的可穿戴设备来说，计算和数据访存是功耗的主要来源。尽管目前已经有研究对 ECG 信号进行压缩处理，但是少有工作将注意力放在 ECG 检测神经网络参数压缩中。由于神经网络使用训练好的权重进行数据处理，同时由于网络的工作机制

等原因，这些权重存在多次复用的特点，因此对神经网络权重进行压缩能有效降低硬件消耗。

尽管所提出的新型生理信号检测算法网络规模相对较小，但是在网络推理阶段，大量高位宽数值计算，仍会导致该网络在进行心律失常识别过程中产生的大量的功耗和内存需求。本设计根据神经网络的特点和心律失常检测场景的特征，采用量化的方式将网络权重量化，可以进一步降低神经网络在使用过程中的硬件开销。

3.3.1 量化原理

量化简而言之就是把网络权值从高精度转化成低精度，例如将 32 位浮点数转化成 8 位定点数或二值化即 1bit，量化可以看作是噪声的一种来源，量化后的模型效果与原来相近，模型准确率等指标与原来相近，模型大小变小，运行速度加快。

量化本质上只是对数值范围的重新调整，可以理解为是一种线性映射。接下来，将以 8bit 整型量化为例，解释量化的原理。

训练好的神经网络由浮点运算构成。浮点数 FP32 和定点数 INT8 的值域分别是$[(2-2^{-23})\times 2^{127},(2^{23}-2)\times 2^{127}]$和$[-128,127]$。取值数量大约分别为$2^{32}$和$2^{8}$。因此，将网络从 FP32 转换为 INT8 并不像数据类型转换截断那样简单，必然存在近似问题。

若我们用 r 表示浮点数，q 表示量化后的定点整数，那么浮点数和定点数的关系可以由式(3.1)和(3.2)得到[34]。

$$r = S(q - Z) \tag{3.1}$$

$$q = round\left(\frac{r}{S} + Z\right) \tag{3.2}$$

其中 S 表示浮点数与定点整数之间的比例关系，Z 表示零点，即浮点数 0 量化之后再定点整数中的位置。他们的计算方式为：

$$S = \frac{r_{max} - r_{min}}{q_{max} - q_{min}} \tag{3.3}$$

$$Z = round\left(q_{max} - \frac{r_{max}}{S}\right) \tag{3.4}$$

在卷积神经网络中，假设输入为 x，然后统计出输入数据各层特征图的最大值和最小值，利用公式(3.3)和(3.4)计算出S_x和Z_x。同样，我们可以得到权重 w 的S_w和Z_w。

这样，卷积运算就可以由式(3.5)变为式(3.6)。

$$y^{i,k} = \sum_{j=1}^{N} x^{i,j} w^{j,k} \tag{3.5}$$

$$q_y^{i,k} = M\sum_{j=1}^{N}\left(q_x^{i,j} - Z_x\right)\left(q_q^{j,k} - Z_w\right) + Z_y \tag{3.6}$$

其中$M = \frac{S_w S_x}{S_y}$。

于是，就可以通过定点数计算完成卷积，最终结合式(3.1)完成反量化，就可以得到最终的结果。

3.3.2 量化实验

对模型进行量化，首先要确定需要量化的全精度模型，本课题中采用的是上文识别准确率位 94.7%的最佳模型。量化工具，选用 Qkeras，采用如图 3-11 所示的训练后量化，即在 32 位浮点数全精度模型训练完成后，对权重进行量化，输入数据与特征图也在推理阶段之前进行量化。

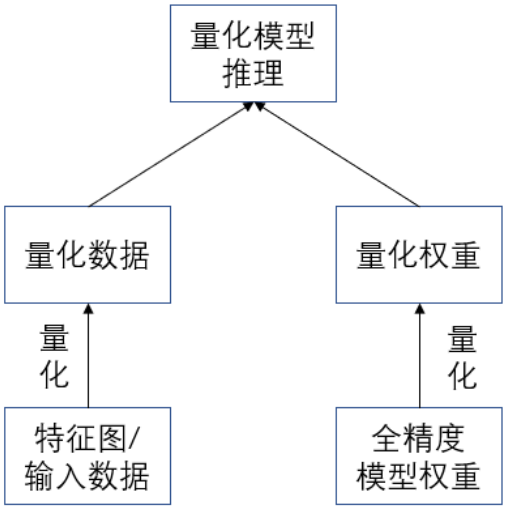


图 3-11 训练后量化

如图 3-12 所示，通过对网络模型权重分布统计和对输入数据以及各层特征图分布的统计之后，发现大多数权重和数据分布在[-2,2]之间，因此将量化之后的定点数范围设置为(-4,4)，即 1 位符号位，2 位整数位，其余为小数位。

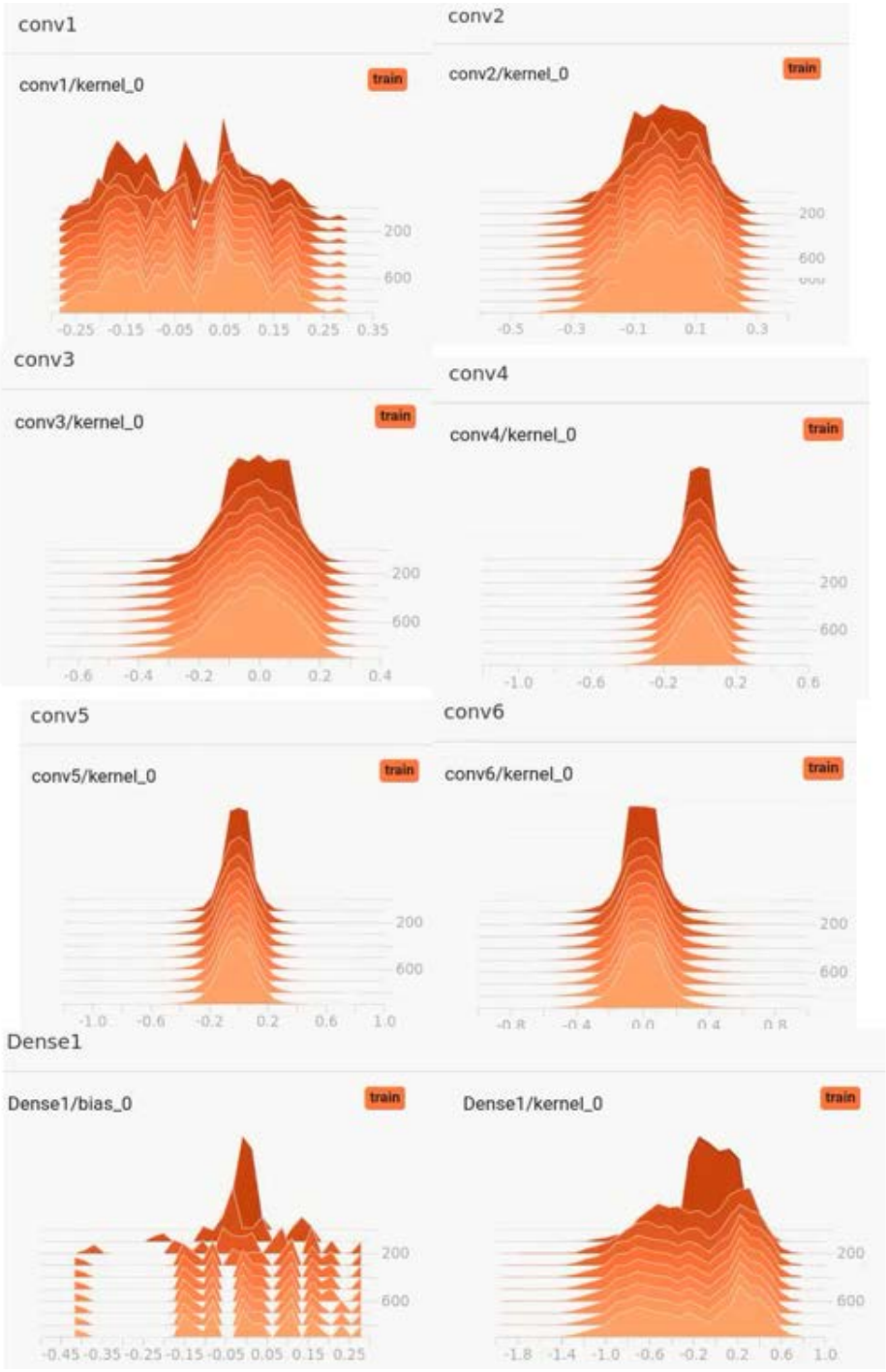


图 3-12 网络模型各层权重分布图

设计量化流程如图 3-13 所示。分别对全精度的最佳模型和第 500 轮训练所得模型进行 16-bit、8-bit、4-bit 量化，要求精度下降不超过 2%。图 3-14 记录了量化过程中，模型识别准确率的变化情况。可以看出，当进行 16-bit 量化时，模型准确率几乎不受到影响；当进行 8-bit 量化时，模型识别准确率略微下降了 0.3%；当进行 4-bit 量化时，模型识别准确率快速下降，只有 88.4%，已不满足本课题中的设计指标要求。因此，最终选择对最佳模型进行 8-bit 量化，定点分布如图 3-15 所示。量化后，模型识别准确率为 94.4%。

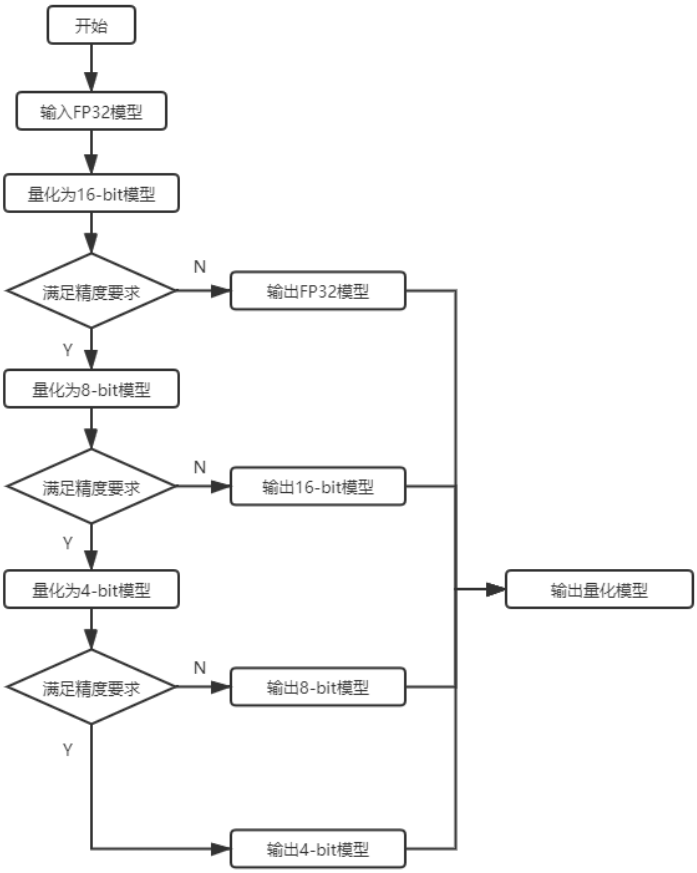


图 3-13 量化流程图

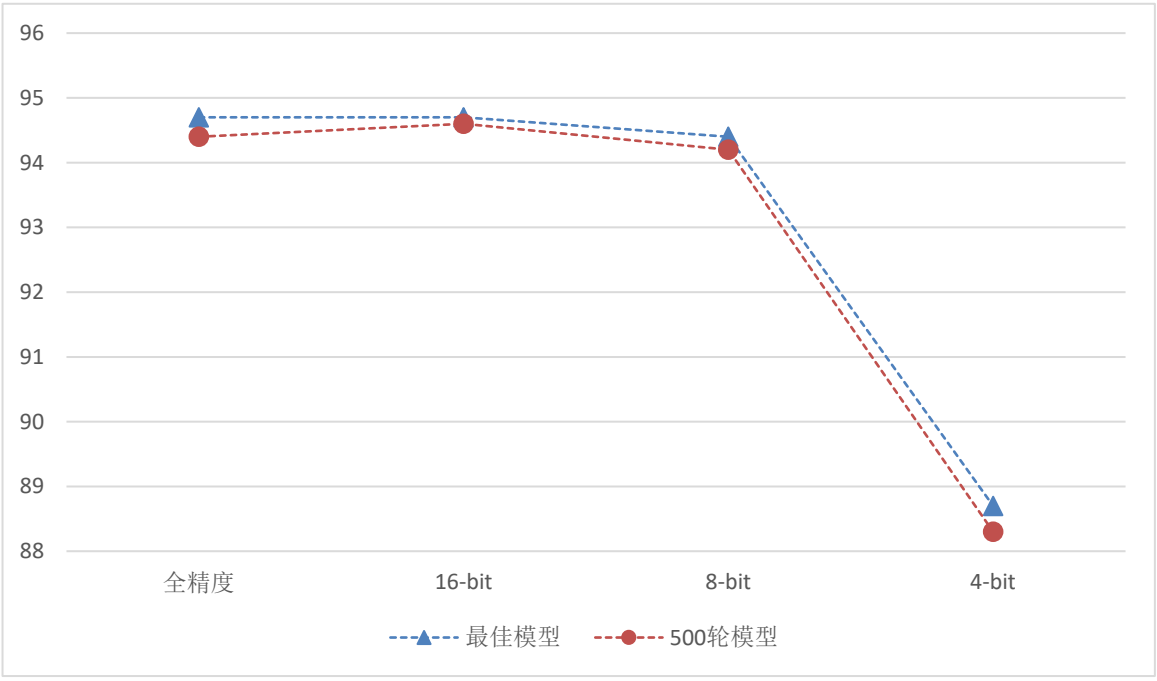


图 3-14 不同位宽量化结果对比

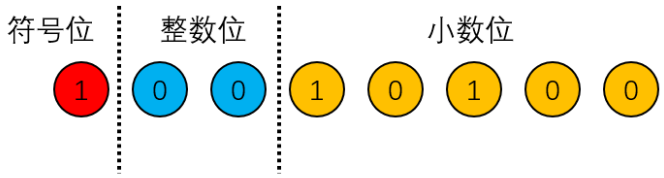


图 3-15 8-bit 量化定点分布(以-0.625 为例)

3.4 本章小结

本章节主要由针对心电信号的基于卷积神经网络的新型生理信号检测算法及算法模型参数压缩两部分组成。在第一部分中，本章提出了一种直连型卷积神经网络结构，并探索了不同网络拓扑结构对于网络性能的影响，并在 MIT-BIH 心律失常数据集上进行了验证，最终实验结果表明，最佳网络模型的识别准确率达到了 94.7%。在第二部分中，本章讲述了量化的理论原理，并对最佳网络模型进行了实验，实验结果显示 8bit 为最合适的量化位宽，量化之后的网络模型识别准确率为 94.4%。

第四章 生理信号检测卷积神经网络硬件加速器电路设计

在第二章的卷积神经网络基本原理以及第三章中所提出的新型生理信号检测算法的基础上，本章设计了一种算法所对应的硬件加速器，首先描述了本设计的总体架构；然后对加速器存储区划和数据流模型进行了详细论述；然后分析了卷积计算的映射方式；最后，描述了加速器控制方式和数据调度原理。

4.1 总体架构设计

如图 4-1 所示，加速器分为 Configurator, Weight Buffer, PE_array, In Out Buffer, Input Regfile, Output Regfile, Relu, Pooling 共 8 个模块，有三股外部的数据流输入，分别是网络参数数据流、权重数据流和 ECG 信号数据流。

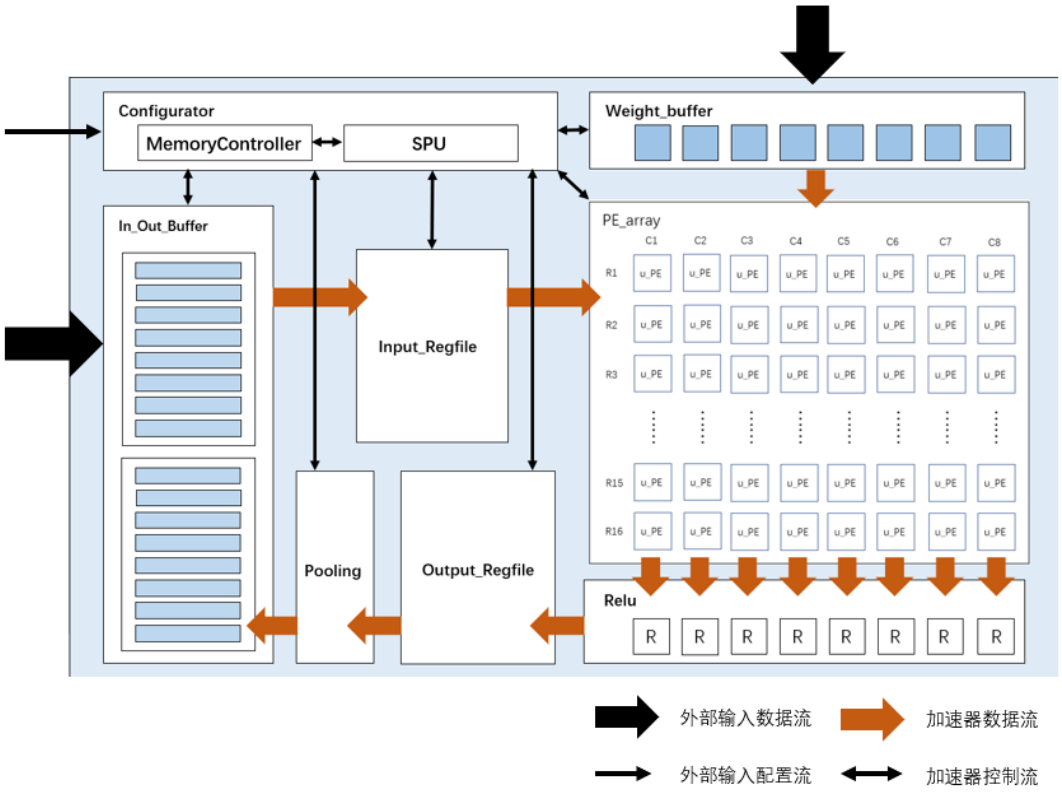


图 4-1 ECG 加速器架构图

当系统启动时，Configurator 中的 Memory controller 模块控制 SPI 总线完成数据调度，将权重和特征图数据载入加速器。ECG 信号输入到第一块 RAM 中，接着按批次输出到 Input Regfile，Input Regfile 把数据按一定的顺序输出到 PE 阵列里与权重进行卷积计算。卷积计算结束，后数据输入 Relu 模块进行激活，之后把完成激活的结果输出到 Output Regfile，

待数据整理好将其发送到 Pooling 模块进行池化操作，最后池化结束将结果输出到 Inout Buffer 的第二块 RAM 中，待所有结果都进入到 RAM 之后，第一层运算结束。第二块 RAM 的数据是下一层的输入，第一块 RAM 即将存放下一层的输出结果，两块 RAM 如此乒乓操作，直到 6 层的卷积层计算结束。

4.2 数据存储及数据流分析

4.2.1 数据存储方案

为了减少在数据读取过程中花费的时间和访存功耗，本设计对数据存储方式进行了优化设计，提出了片上缓存、中间缓存、寄存器堆的三级存储方案。

图 4-2 显示了加速器的整体存储划分，其中片上缓存包括 In Out Buffer 和 Weight Buffer 两个模块，分别存储特征图数据和权重数据。In Out Buffer 有两块 RAM，每块 RAM 里面有个 8 个 BRAM，每个 BRAM 的深度是 1024，位宽是 8bits，每个 BRAM 为 1KB 大小。Weight Buffer 有 8 个 BRAM，每个 BRAM 的深度是 2048，位宽也是 8bits，每个 BRAM 为 2KB 大小。

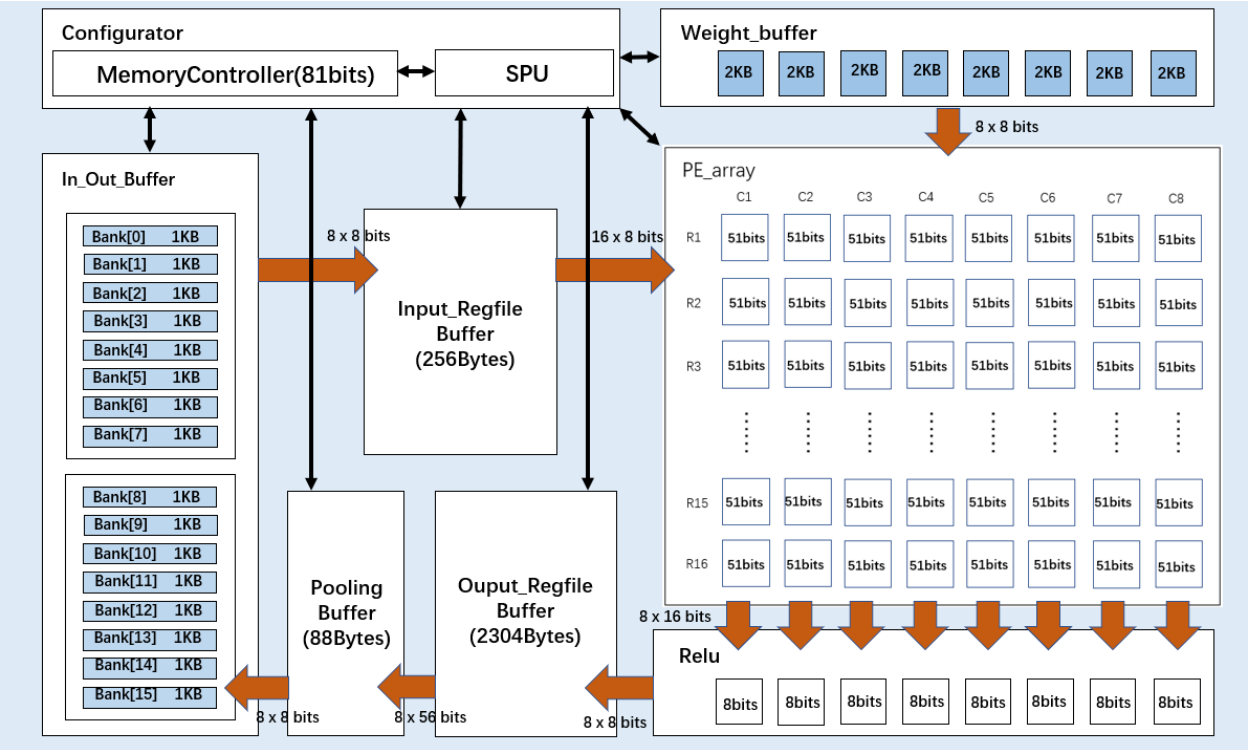


图 4-2 ECG 加速器的存储划分

中间缓存包括 Input Regfile 和 Output Regfile 两个模块，Input Regfile 模块用于缓存脉

动阵列在下一次计算中所使用到的所有输入数据，起到数据复用的功能。由于卷积核大小 K>卷积核步长 S，在脉动阵列的相邻两次计算过程中的输入数据存在重叠，使用 Input Regfile 模块可以避免对 In Out Buffer 中的同一个地址多次访问的现象。Input Regfile 模块原理如图 4-3 所示。

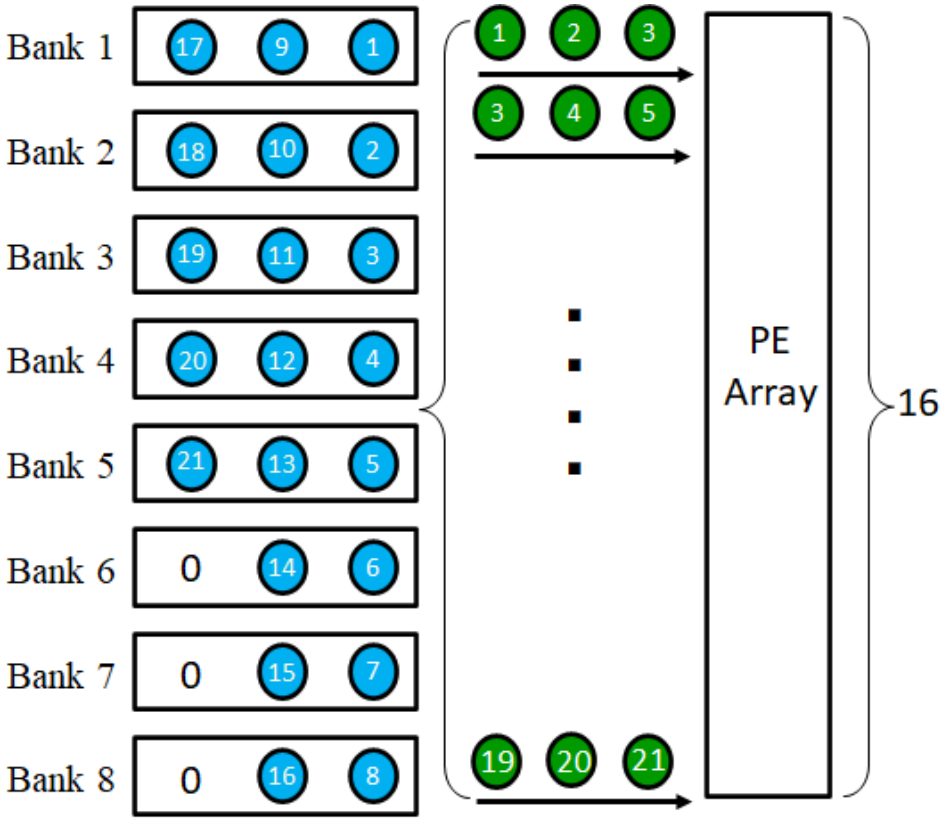


图 4-3 Input Regfile 模块原理图

Input Regfile 存放 Hu × 1 的输入数据，其中 Hu 为脉动阵列在一次计算中所涉及的输入数据数量(Hu)与当前层的卷积核大小 K，卷积核步长 S 和 PE Array 的 Map 输入端口数量 R 有关，其计算公式如下：

$$Hu = (R - 1) \times S + K \tag{4.3}$$

本设计中，各网络层对应的 Input Regfile 大小如下表 4.3，最终确定 256Byte。

表 4.3 每层网络 Input Regfile 所需大小表

网络层索引	Hu	对应 Input Regfile 大小
1(Conv1)	46	46 Byte
2(Conv2)	42	42 Byte
3(Conv3)	23	23 Byte

续表 4.3 每层网络 Input Regfile 所需大小表

网络层索引	Hu	对应 Input Regfile 大小
4(Conv4)	19	19 Byte
5(Conv5)	18	18 Byte
6(Conv6)	256	256 Byte(max)
8(Dense)	72	72 Byte

中间缓存的另一组成模块是 Output Regfile 模块，用于实现输出数据复用的功能。脉动阵列每一列通道一次输出 16 个计算结果，有 8 列通道，因此每一列通道对应有一个 Output Regfile。由于当前输出的计算结果与上一次输出的计算结果还有一定池化联系，所以将一个 Output Regfile 的大小定为 32 byte，一次卷积计算的结果为 16byte，可以存两次卷积计算的结果。低地址 16byte 写入奇数次的计算结果，高地址 16byte 写入偶数次的计算结果，轮流写入即可。

寄存器堆包括 PE 单元内存储中间计算结果的寄存器等。单个 PE 单元内包含 40bits 的数据存储寄存器和 11bits 的控制信号存储寄存器。

4.2.2 数据通路

数据通路为除控制模块外所有模块组成的环路，环路根据卷积神经网络的基础结构，即卷积层、激活函数、池化层的顺序流程进行设计。本课题中基于三级存储方案的数据流模型如图 4-4 所示。

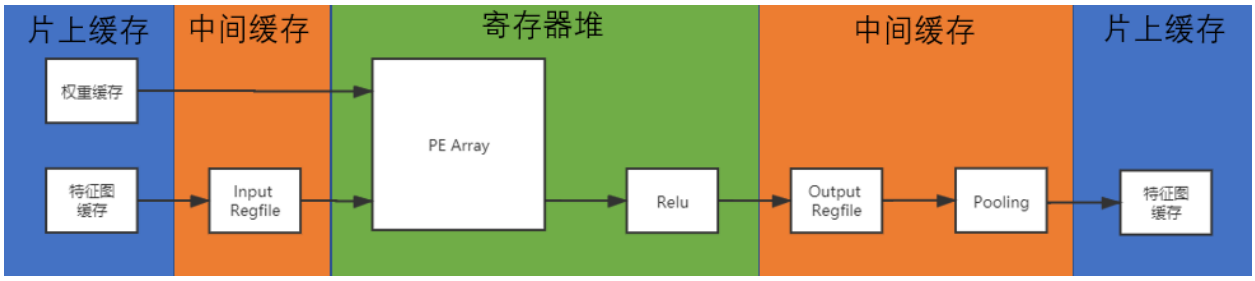


图 4-4 基于三级存储方案的加速器数据流模型

数据通路依照卷积神经网络操作执行顺序设计，从权重/特征图至 Input Regfile 至脉动阵列至激活函数模块至 Output Regfile 至池化模块，最后保存输出的特征图数据。存储区划变化为片上缓存至中间缓存至寄存器堆再至中间缓存，最终回到片上缓存。呈现 1 级 2 级

3 级 2 级 1 级的循环变化模式。存储器间的数据通信根据计算通路进行设计，包括特征图输入输出，权重输出，寄存器堆缓存等。

4.3 计算引擎设计

计算引擎是整个加速器的核心模块。本课题中采用脉动阵列来作为计算引擎。脉动阵列，又称为 PE 阵列，适用于卷积等乘加操作，是目前较为常用的一种计算引擎架构。

4.3.1 设计空间探索

由于 CNN 的网络模型特点，每一层的计算方式都类似，因此设计和优化适合 CNN 网络模型和卷积核大小的脉动阵列对于提高加速器性能具有重要意义。

为了实现最佳的加速器性能，需要对脉动阵列的大小进行设计空间探索。一般来说，脉动阵列越小，数据的复用程度就越高，但同时需要更多的计算次数。因此，有必要根据卷积神经网络模型，在这些限制下探索最佳脉动阵列大小，以最大限度地提高加速器的性能[35]。

脉动阵列利用率、吞吐率和能效是评价一个卷积神经网络硬件加速器的重要指标。其中，利用率是指平均参与计算的 PE 单元数量占总数量的占比。表达式如式(4.1)所示，其中N_{cycle}为计算轮数，R 为阵列行数，C 为阵列列数，N_{cal}为单次计算量。

$$Utilization = \frac{\sum N_{cal}}{N_{cycle} \times R \times C} \tag{4.1}$$

吞吐率被定义为每秒进行的操作数量。表达式如式(4.2)所示，其中N_{ops}为操作数量，T 为计算时间。

$$Throughput = \frac{N_{ops}}{T} \tag{4.2}$$

能效被定义为每秒每瓦进行的操作数量。表达式如式(4.3)所示，其中 E 为总能耗。

$$Power\ Efficiency = \frac{Throughput}{Power} = \frac{\left(\frac{N_{ops}}{T}\right)}{\left(\frac{E}{T}\right)} = \frac{N_{ops}}{E} \tag{4.3}$$

本课题在设计空间探索中，主要分析了不同阵列行列数对以上三个性能指标的影响。实验结果如图 4-5 所示，综合三项指标，最终确定脉动阵列设计参数为：R = 16, C = 8。

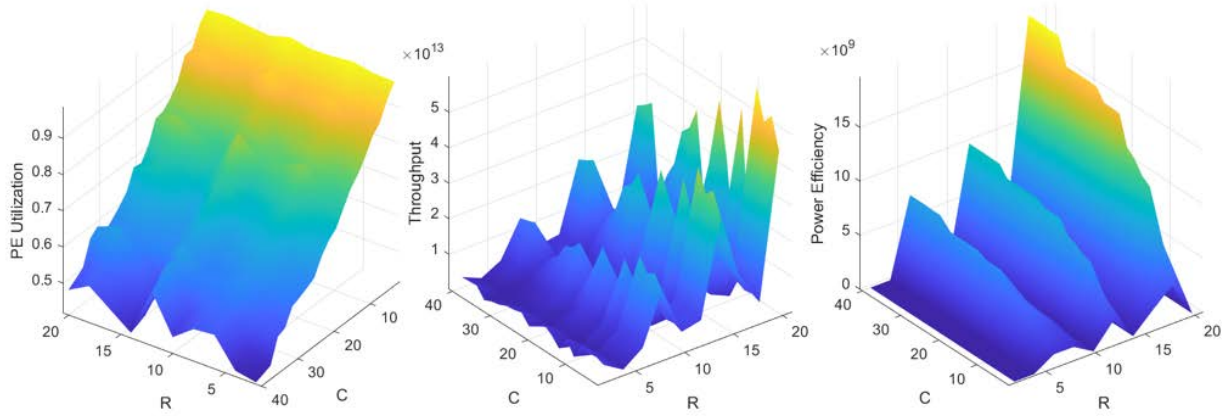


图 4-5 脉动阵列利用率、吞吐率和能效在 R、C 变化下的分布

4.3.2 脉动阵列工作原理

脉动阵列，又称为 PE 阵列，本课题中所设计的脉动阵列大小是 16*8，16 行，8 列，8 列脉动阵列的权重输入对应着 Weight Buffer 的 8 个 BRAM 的输出，16 行脉动阵列的特征图输入对应着 Input Regfile 的输出。

脉动阵列一共含有 128 个 PE 单元，虽然功能比较单一但是为加速器的核心模块，只进行卷积计算，并输出结果。脉动阵列的工作原理如下：在第一个周期的时候，Weight Buffer 第一个 BRAM 和 Input Regfile 输出权重和特征图至第一列 16 个 PE 单元之中，16 个 PE 单元同时进行第一次计算；第二个周期的时候，权重和特征图输入第二列 PE 单元，进行第 2 列 PE 单元的第一次计算，第一列 PE 单元进行第二次计算，如此从左至右地脉动；直到第八个周期，第八列 PE 单元进行第一次计算，第一列 PE 进行第八次计算，整个脉动阵列被激活起来，都在进行着计算工作。当计算完成后，脉动阵列以列的形式输出计算结果，每一行的 PE 单元将计算结果依次传给上一行，最终在第一行的 PE 单元输出。脉动阵列数据流模型如图 4-6 所示。

当收到输入特征图有效信号和输入权重有效信号，并且计算状态机处于进行卷积计算的状态时，脉动阵列开始计算。主要的配置参数包括每次计算需要的周期数，当前计算层数等。脉动阵列计算完成后，将 PE_end 信号置为 1，输出计算结果。

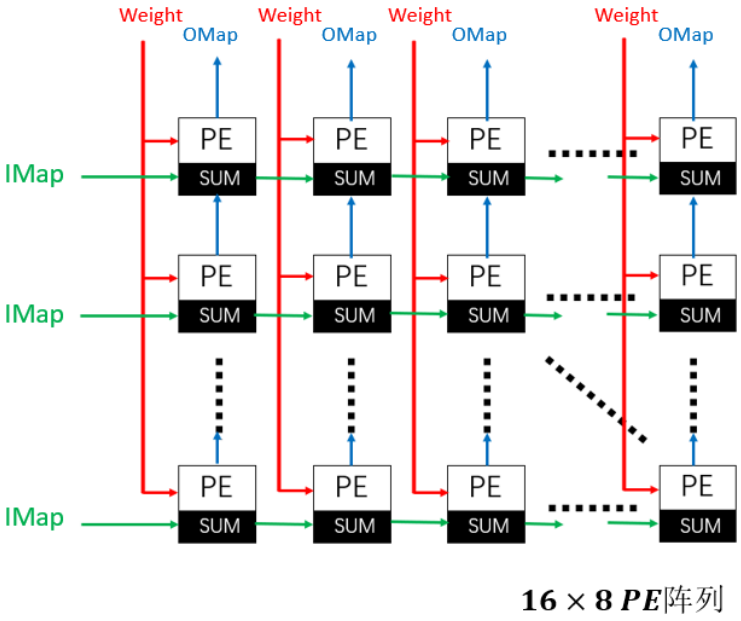


图 4-6 PE 阵列数据流模型

每一个 PE 单元主要由 8bit 的乘法器和 16bit 的加法器，缓存中间值的寄存器、计算执行周期的计数器以及其他基本元件组成。作用包括传递输入数据、权重和输出数据，执行乘累加运算，缓存中间计算结果。当权重和数据使能信号同时置 1 时，执行乘累加运算，并沿行方向传递数据，当输出使能置 1 时开始沿列方向传递输出数据。PE 单元内部结构如图 4-7 所示。

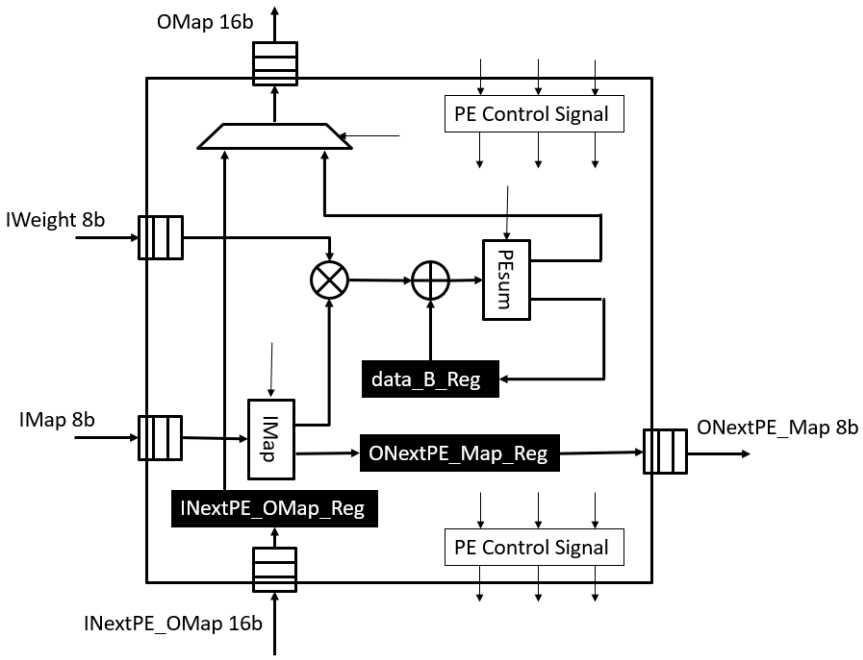


图 4-7　PE 单元内部结构

4.4 主控制器设计与数据调度

主控制器主要包括 State Processing Unit(SPU)子模块和 MemoryController(MC)子模块。SPU 模块主要控制整个系统状态转换。分别有两个状态机：主状态机和计算状态机。MC 模块主要是控制并实现加速器与外部信号的数据交换，数据包括网络层配置参数，网络层权重，ECG 输入信号。交换的数据类型由 SPU 模块状态机决定。

(1) SPU 主状态机：对加速器实现一次推理的控制。状态机状态转换图 4-8 所示。状态如表 4.2 所示。

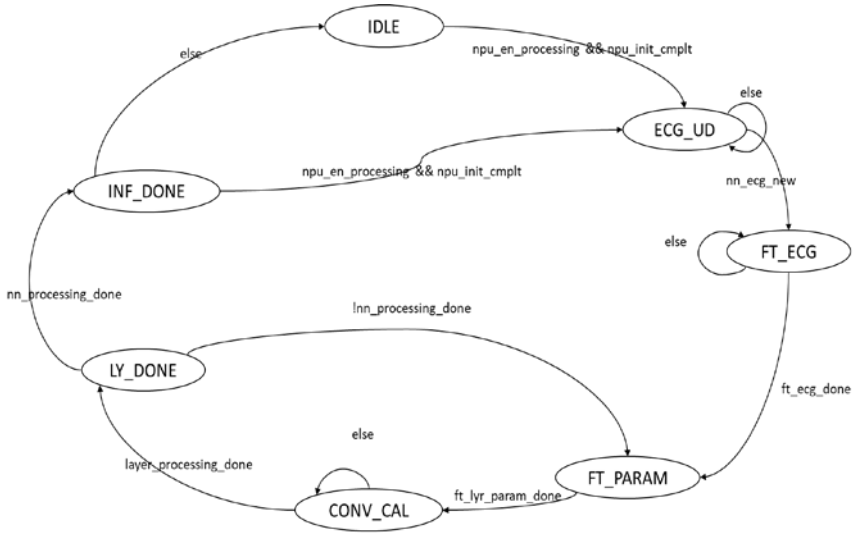


图 4-8 加速器主状态机原理图

表 4.2 主状态机状态表

标志	状态
IDLE	空闲状态。
ECG_UD	等待 ECG 信号采集芯片通过 SPI 接口将 ECG 信号保存在片外存储器中。
FT_ECG	将保存在片外的 ECG 信号通过 Memory Controller 从片外存储器载入到 In Out Buffer 中。
FT_PARAM	从片外存储器(或片内存储器)中载入当前层的网络配置参数，并将其输出给各计算模块和存储模块。
CONV_CAL	卷积计算状态，控制对加速器进行当前层的数据读写和计算。
LY_DONE	当前层计算完毕。
INF_DONE	所有层计算完毕的状态。

(2) SPU 计算状态机：实现单次计算的控制。状态机状态转换图 4-9 所示。状态如表 4.3 所示。

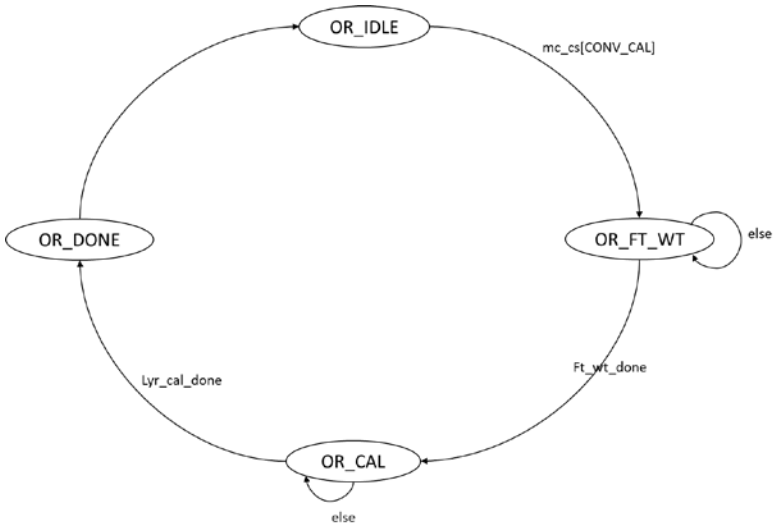


图 4-9 加速器计算状态机原理图

表 4.3 计算状态机状态表

标志	状态
OR_IDLE	空闲状态。
OR_FT_WT	加载 weight 值的状态，并判断 ft_wt_done 是否置 1，若置 1 则表示权值加载完毕，进入计算状态，否则继续保持 OR_FT_WT 状态。
OR_CAL	进行卷积计算的状态，并判断 lyt_cal_done 是否置 1，若置 1 则表示卷积计算完毕，进入计算结束状态，否则继续保持 OR_CAL 状态。
OR_DONE	卷积计算完毕的状态，回到空闲状态。

加速器的数据调度方式由 SPU 和 MC 模块协同决定，流程图如图 4-10 所示，进行卷积时，首先读取当前层的参数，然后等待权重和特征图有效信号，加载权重和特征图进行计算。执行计算时，读取数据输入 PE 阵列，PE 阵列每轮计算一定数量的输入点，同时进行计数以判断输入 map 是否读取完成。PE 阵列输出数据时，每列输出 16 个数据进行后续处理最终存入片外存储器中，同时进行计数以判断输出是否完全写入。

当输入输出的计数满足要求时，当前层计算完成，进入下一层。

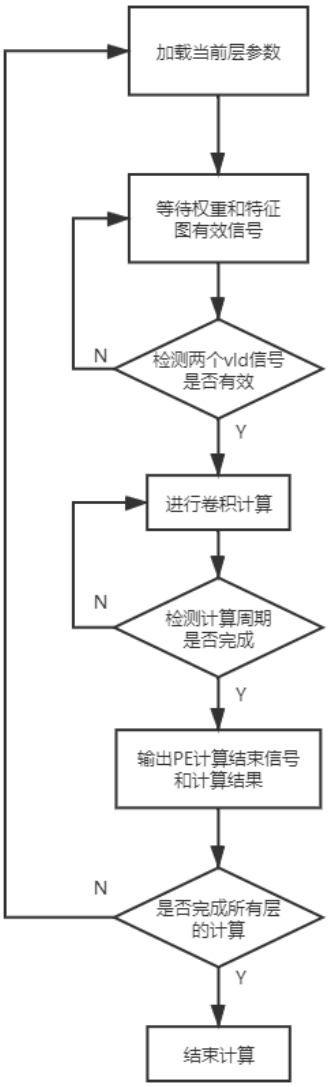


图 4-10 加速器数据调度流程图

4.5 本章小结

本章设计了一种基于卷积神经网络的新型生理信号检测算法硬件加速器电路，并针对访存功耗、存储功耗和计算功耗进行了优化设计。首先，提出了以三级存储结构为基础的数据存储方案，并分析了对应的数据流模型；然后，详细描述了计算引擎的设计方法，包含脉动阵列设计空间探索，以及工作原理和 PE 单元内部结构；最后，论述了加速器的控制方法，并分析了一层卷积层计算的数据调度模型。

第五章 仿真与测试

为了正确的测试以及评估本课题提出的基于卷积神经网络的生理信号检测算法硬件加速器电路，本章节对设计好的加速器各个模块进行独立测试和模块间共同测试从而实现对加速器各个模块的功能的验证和加速器整体功能的验证；同时，结合算法模型，对加速器的运行速度、资源使用和功耗进行评估；最终，对硬件加速器的识别准确率进行实验，并与其他平台进行纵向和横向的对比。

5.1 功能仿真

为了对本课题所设计的基于卷积神经网络的新型生理信号检测算法硬件加速器电路进行有效的功能仿真，本课题建立了以 Verilog 硬件语言为基础的 test bench 仿真测试文件。下面进行各功能模块的仿真结果展示。

5.1.1 控制模块仿真测试

Configurator 里面包含 SPU 和 MemoryController，主要是测试 SPU 和 MemoryController 之间的通讯情况，检测状态机的跳转状态即可。从图 5-1 可以看出 nn_layer_cnt 从 1 到 6 再回到 1，即 6 层网络的状态全部跑通，且主状态机和计算状态机跳转正常，完成了对该功能的测试。

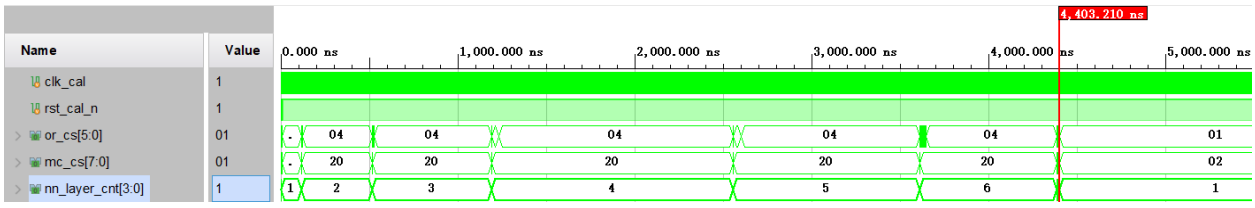


图 5-1 控制功能仿真测试结果

5.1.2 数据输入联合仿真结果

这部分功能模块仿真测试主要涉及 In Out Buffer 和 Input Regfile 两个模块。In Out Buffer 模块获取各项网络参数以及数据读写的地址后，状态机跳转，开始执行数据读出。Input Regfile 计算当前所读取的地址之后，接收 In Out Buffer 的数据并重新排列数据，在当前读取操作的次数满足之后输出至脉动阵列。具体测试方法如下：

(1) In Out Buffer 模块收到脉动阵列计算结束信号时开始进行数据传输，通过观察

数据输出所需的次数和当前计数器变化，可以完成验证。

(2) Input Regfile 模块根据 Bm_cnt 计算写入的首地址 wr_addr，一次写入 In Out Buffer 的 Bank 数量（8 个）的数据。通过观察地址变化，可以完成验证。

(3) Input Regfile 模块接收脉动阵列计算结束信号时，每个周期写地址加 1 计算输出的首地址 rd_addr，控制数据输出次数。一次写入脉动阵列的 R 数量(16 个)的数据。

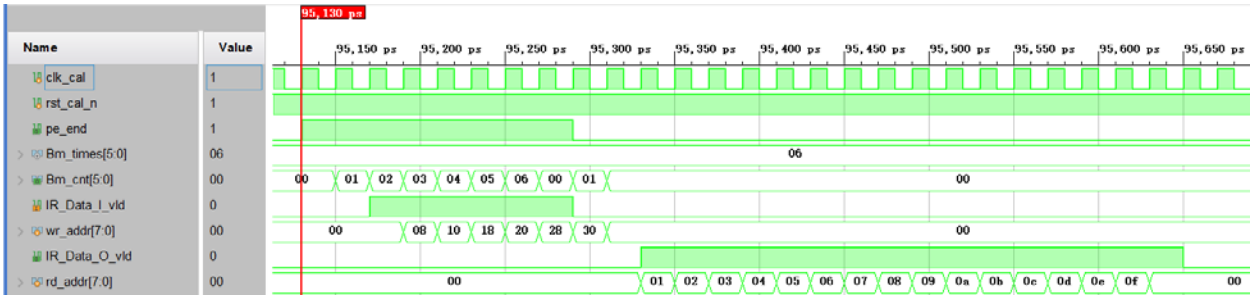


图 5-2 数据输入联合仿真测试结果

数据输入功能 In Out Buffer 和 Input Regfile 模块联合仿真测试结果如图 5-2 所示，数据输出此时和读写地址变化正常，功能完全正确。

5.1.4 权重读写模块仿真测试

这部分功能模块仿真测试主要涉及 Weight Buffer 模块。主要控制权重数据的读写。具体测试方法如下：

(1) Weight Buffer 写入数据时每一个权重的位宽是 1byte，每一次同时写入 8 个数据分别写入 8bank 中，因此对于第一层来说只需要写入 16clk 就可以把第一层的权重值全部写入到 bank 中。观察 16 个时钟内的地址变化情况可以验证写入功能。

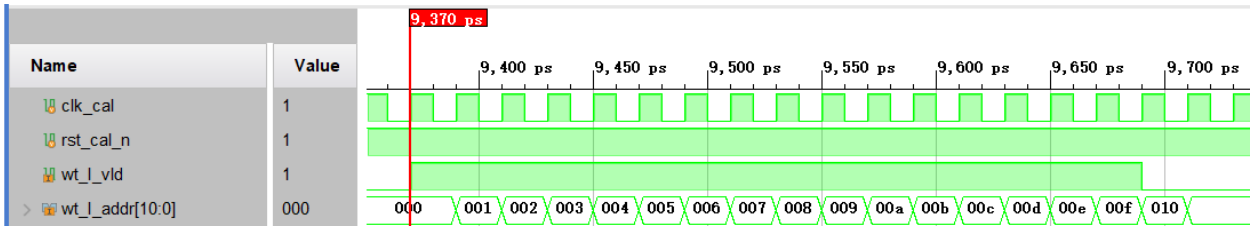


图 5-3 Weight Buffer 模块写入测试结果图

(2) Weight Buffer 接收到 MC 模块的控制后信号，开始读出 bank0 中的数据。通过观察波形，看到每个 bank 间隔一个时钟依次读出数据，与 PE 阵列从左到右的脉动计算匹配。可以完成功能验证。

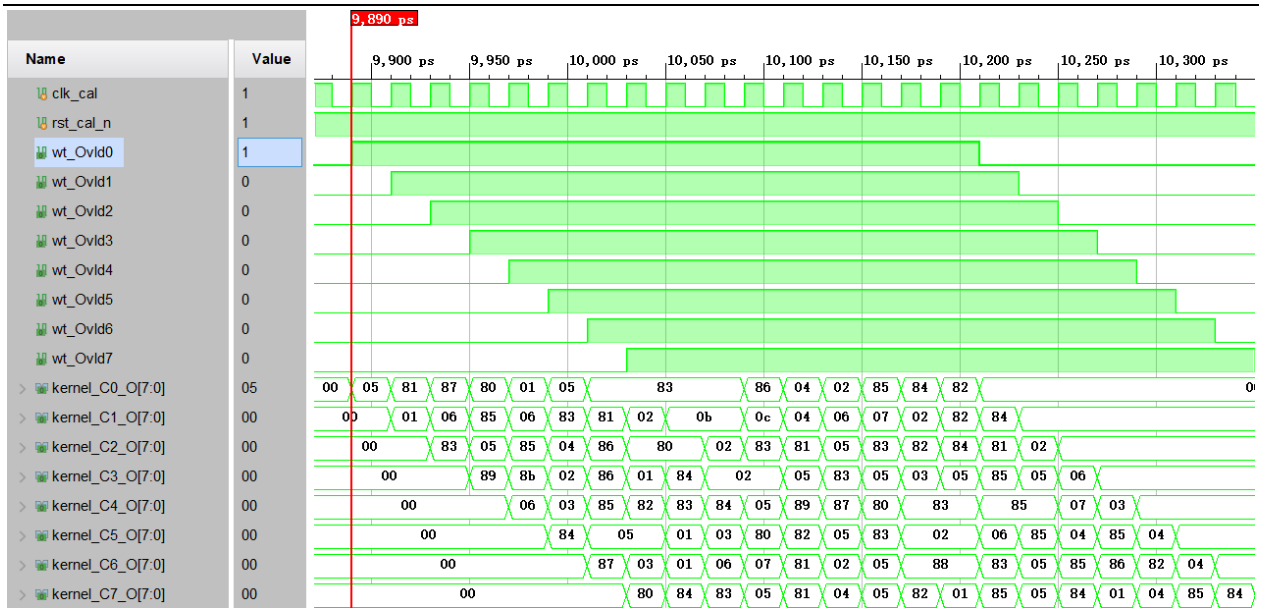


图 5-4 Weight Buffer 模块读出测试结果图

权重读写模块读写仿真测试结果如图 5-3 和图 5-4 所示，数据写入时的地址变化正常，数据读出时，输出顺序与脉动阵列匹配，功能完全正确。

5.1.5 计算功能仿真测试

PE 单元仿真结果如图 5-5 所示，主要验证 PE 单元的计算结果正确性。Multi 是乘法器的运算结果，刚开始两个输入数据为 05 和 0e（HEX），即乘法结果为(5)*(14) = 70，即 16 进制的 0046 结果正确。紧接着的两个周期 81*11 和 87*13，81*11 即为(-1)*(17) = -17，87*13 即为(-7)*(19) = -133，所以 Multi 的结果输出是 8011 和 8085（十六进制的-17 和-133）。PEsum 是加法器的运算结果，第一个周期就是乘法器的结果 0046，第二个周期的 PEsum 结果为 0035 是 0046 和 8011 相加，结果正确。

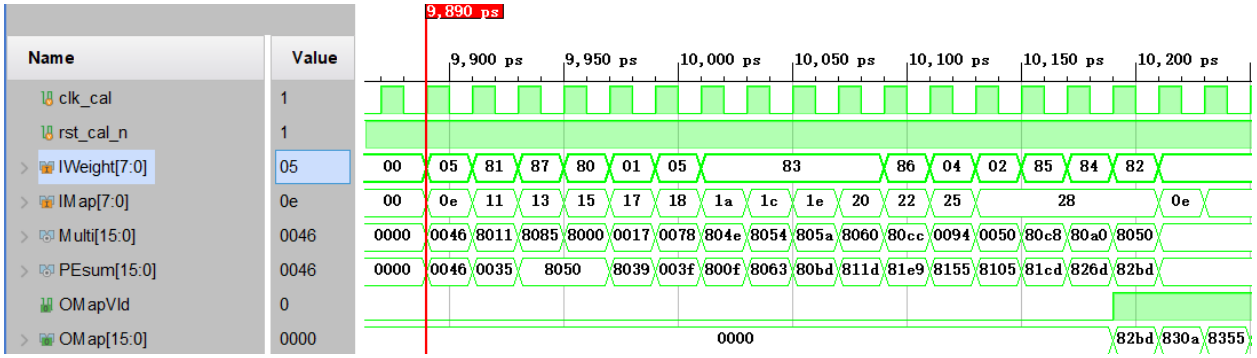


图 5-5　PE 模块计算测试结果图

5.1.6 数据输出联合仿真测试

这部分功能涉及了 Relu8，Pooling8，Output Regfile，Pool2IOB 四个模块。每一列的 PE 都接一个 Relu，Output Regfile 的数据输入是 Relu8 的计算结果输出，再输入 Pooling8 和 Pool2IOB 之中，最终重组后进入 In Out Buffer。将重点验证 Output Regfile 模块功能，具体验证方法如下：

(1) Output Regfile 数据写入时，单数次的 16 个 byte 数据写入地址应该为 01 至 0f，双数次的 16 个 byte 数据写入地址应该为 11 至 1f，通过观察写入地址变化，可以完成验证。

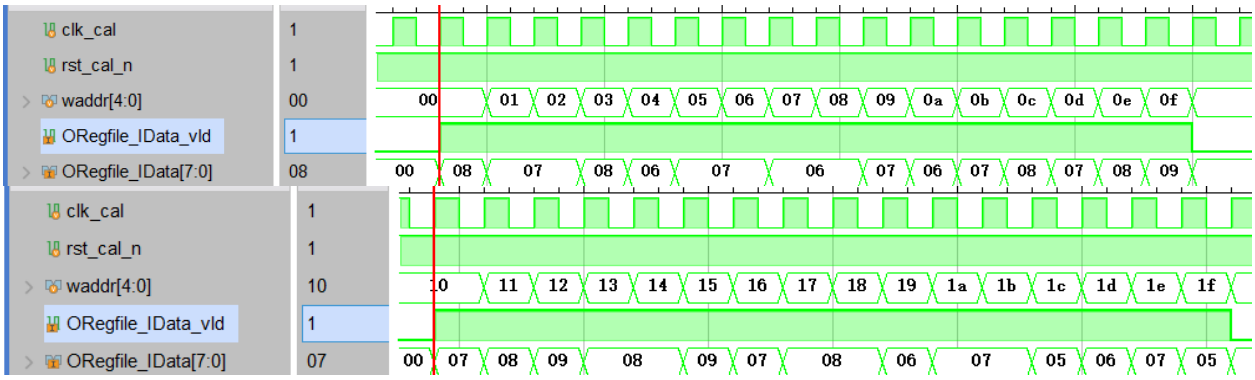


图 5-6 Output Regfile 模块写入测试结果图

(2) Output Regfile 数据读出时，假设池化层步长为 2，最大化池大小为 2，地址从 0 开始读，到地址 e 结束，每次读出 2 个 byte 的数据输出给 RP，通过观察数据的间隔为，可以验证功能是否正确。

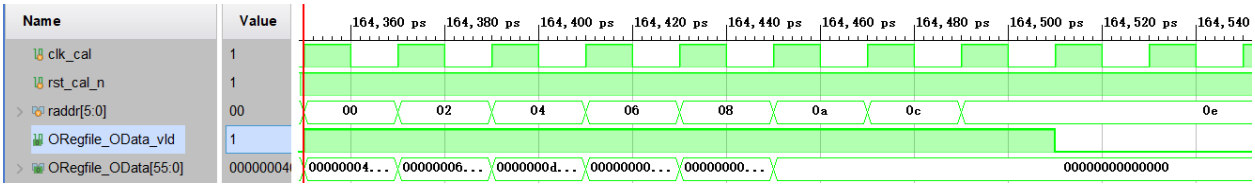


图 5-7 Output Regfile 模块读出测试结果图

数据输出联合仿真测试结果如图 5-6 和图 5-7 所示，数据写入时，奇数次和偶数次地址变化正确，分别为 01 至 0f 和 11 至 1f，地址变化正确。数据读出时，地址从 00 至 0e 结束，每次读出 2 个 byte 的数据输出给 RP，数据的间隔为 2 对应步长为 2，验证结果正确。

5.2 性能评估

本课题中的性能评估是基于 Xilinx 公司的 Artix-7 系列的 FPGA xc7z020clg484-1。该

FPGA 包含 220 个用于实现乘法、加法、乘累加和逻辑运算的数字信号处理器（Digital Signal Processing，DSP），140 个容量为 36Kb 的块随机存取存储器（Block RAM，BRAM），53200 个查找表（Look-up tables，LUT），200 个 I/O 接口，106400 个触发器（Flip-flops，FF）等资源。

软件计算是使用 CPU 进行计算的，CPU 计算能力往往有限，但具有灵活性、普遍性。硬件计算是利用专用硬件模块进行计算，计算效率较高，具有专业性。基本上，现在大多数机器学习都是基于神经网络，但需要大量的计算，这些计算不够快的 CPU，所以使用硬件加速器的设计，因为有大量的计算设备可以做矩阵计算，如中国科学院的 DianNao，谷歌的 TPU 等，致力于神经网络加速计算。本设计中的硬件的仿真时间如图 5-8 所示。

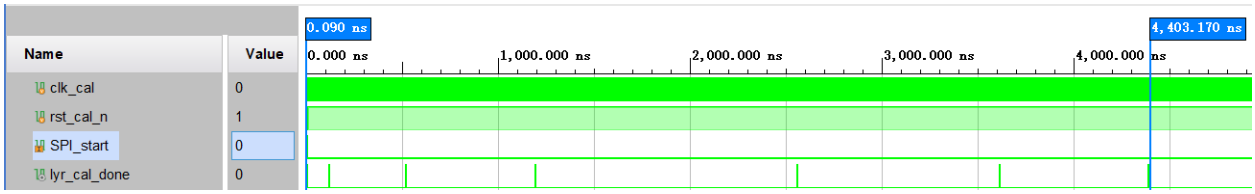


图 5-8 硬件仿真时间

从写入指令的 90 时间节点截止到 CNN 加速器到第 8 层结束的 4,403,170 时间节点，仿真周期是 20ps,所以硬件加速器计算总共用了（4403170-90）/20=220154 个周期，若是算时间的话，220154 个周期*20ps=4.40308μs（约等于 4.4μs）。

在 Vivado2018 上对于基于 ECG 信号的高能效智能心率检测 SOC 进行性能测试，得到的资源利用情况如图 5-9 所示。FPGA 资源使用表如表 5.1 所示。所统计的均为加速器模块内部电路所占用的资源。

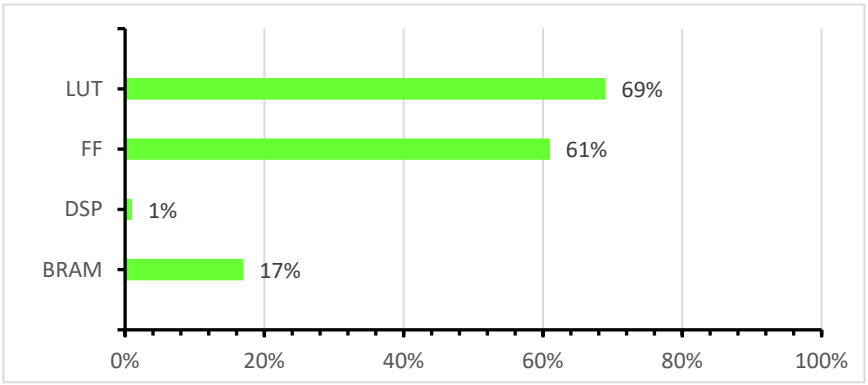


图 5-9 FPGA 资源使用情况图

表 5.1 FPGA 资源使用情况表

Resource	Utilization	Available	Utilization%
LUT	36778	53200	69.13%
FF	64855	106400	60.95%
DSP	1	220	0.45%
BRAM	24	140	17.14%

完成量化之后，参照 moons 的工作[36]建立网络层的能耗模型，从式(5.1)至(5.4)可以看出功耗由三部分组成，分别为计算功耗，权重访存功耗和激活值访存功耗。

$$E_C = E_{MAC} \times (N_c + 3 \times A_s) \tag{5.1}$$

$$E_W = E_M \times N_s + \frac{E_L \times N_c}{\sqrt{p}} \tag{5.2}$$

$$E_A = 2 \times E_M \times A_s + \frac{E_L \times N_c}{\sqrt{p}} \tag{5.3}$$

$$E_{HW} = E_C + E_W + E_A \tag{5.4}$$

其中，E_{MAC}、E_M、E_L分别为单次乘加操作能量消耗、单个数据主 SRAM 存储能量消耗和本地 SRAM 存储能量消耗。这些都可以使用含量化位宽 q 的表达式表示，单位均为 pJ。

$$E_{MAC} = 5 \times \left(\frac{q}{16}\right)^{1.25} \tag{5.5}$$

$$E_M = 10 \times \left(\frac{q}{16}\right) \tag{5.6}$$

$$E_L = 5 \times \left(\frac{q}{16}\right) \tag{5.7}$$

最终，代入乘加操作数量N_C、激活函数A_S、参数权重数量N_S后，将数据复用程度 p 设为 1，最终得到估算的单次识别硬件总能耗E_{HW}为 19.93μJ。

图 5-10 为生理信号检测卷积神经网络硬件加速器电路的 FPGA 功耗综合评估情况。处理器工作频率为 667MHz，去除 PS7 处理器功耗后，加速器电路总功耗仅为 0.651W。其中静态功耗为 0.221W，动态功耗为 0.43W。结合单次运行时间 4.4μs，可得单次识别能耗为 2.86μJ，与估算的能耗数量级接近，实验结果较为合理。

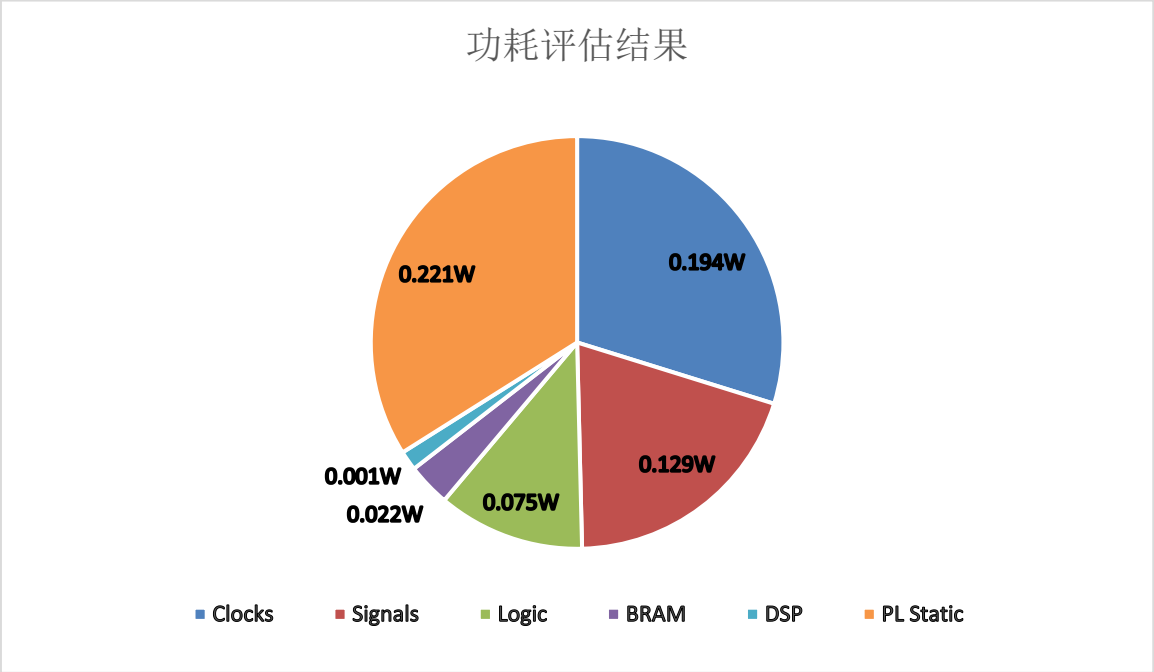


图 5-10 FPGA 功耗综合评估图

5.3 加速器准确率评估

在将算法映射为电路时，由于数据溢出或位宽截取等问题，加速器的识别准确率必然受到影响，相比软件算法的识别准确率出现下降。为了评估加速器的识别准确率，并缩短实验时间，将选取部分 ECG 信号输入数据进行估算。

首先，将加速器精度 ACC 定义为：

$$ACC = \frac{T_h}{N} \tag{5.8}$$

式(5.8)中T_h和 N分别为加速器预测正确个数和测试集总个数。

将软件算法精度 acc 定义为：

$$acc = \frac{T_s}{N} \tag{5.9}$$

式(5.9)中T_s和 N分别为软件算法预测正确个数和测试集总个数。

将算法到电路的映射能力称之为一致率 CR：

$$CR = \frac{T_h}{T_s} \tag{5.10}$$

则，可以得到

$$ACC = CR \times acc \tag{5.11}$$

此时，选取召回率前三的类别，每个类别选取数据集总的 3%，一共选 20 个。由于，这五类召回率高，可以近似认为选取的 3%的数据，都被算法预测正确。所以 CR 可以近似为：

$$\mathrm{CR} = \frac{\mathrm{T_h}}{\mathrm{T_s}} \approx \frac{\mathrm{T_c}}{\mathrm{T_s}} \tag{5.12}$$

最终，基于式(5.11)与(5.12)，可以只利用少量 ECG 信号输入数据，来对加速器识别准确率做出评估，更科学的衡量本课题所提出的生理信号卷积神经网络硬件加速器电路在 MIT-BIH 数据集上的表现。加速器识别准确率测试表格如表 5.2 所示。Recall 召回率前三的类别分别为 1 NSR 7 PVC 14 LBBBB。

表 5.2 加速器识别准确率测试表格

	数据集总数	测试集数量	加速器测试数量	仿真测试正确数量
1 NSR	283	85	10	9
2 APB	66	20	-	-
3 AFL	20	6	-	-
4 AFIB	135	41	-	-
5 SVTA	13	4	-	-
6 WPW	21	7	-	-
7 PVC	133	40	5	5
8 Bigeminy	55	17	-	-
9 Trigeminy	13	4	-	-
10 VT	10	3	-	-
11 IVR	10	3	-	-
12 VFL	10	3	-	-
13 Fusion	11	4	-	-
14 LBBBB	103	31	5	5
15 RBBBB	62	19	-	-
16 SDHB	10	3	-	-
17 PR	45	14	-	-

最终，根据表 5-4 的结果与式(5.11)与(5.12)，得到加速器识别准确率 ACC 为 89.7%

$$\mathrm{ACC} = \mathrm{CR} \times \mathrm{acc} = \frac{19}{20} \times 94.4\% = 89.7\% \tag{5.13}$$

5.4 实验结果分析

本课题提出了一种新型生理信号检测算法，对卷积神经网络算法的结构进行了优化，并对模型参数进行了压缩，使其有助于低功耗加速器硬件电路的实现。结合减少访存功耗

与脉动矩阵的设计思路，完成了对应的加速器电路的设计。最终的实验数据包括 FPGA 逻辑资源消耗、在 MIT-BIH 数据集上的识别准确率以及运行速度，并与 CPU、GPU 上的运行速度进行对比，如表 5.3 所示。

表 5.3 新型生理信号检测算法不同平台实验结果对比

实验数据	FPGA Xilinx xc7z020clg484-1	CPU Intel Core(TM) i7-9750H	GPU NVIDIA GeForce RTX 2060
LUTs 使用情况	36778	-	-
FFs 使用情况	64855	-	-
DSPs 使用情况	1	-	-
BRAM	24	-	-
识别准确率	89.7%	94.7%	94.7%
单次识别运行时间	4.4μs	186.87μs	24.66μs

从表 5.3 中可以看出，加速器识别准确率为 89.7%，相比 CPU 和 GPU 平台 94.7%的识别准确率下降了 4.7%，同时，FPGA 上单次识别运行时长只有 4μs，速度是 CPU 平台的 42 倍，是 GPU 平台的 6 倍。

本课题所设计的基于卷积神经网络的新型生理信号检测算法以及硬件加速器电路的设计指标与相关性能指标的实验结果对比如表 5.4 所示。主要分为以下三点：

(1) 优化模型，模型大小压缩率达到 25%。完成设计指标要求。

(2) 设计的生理信号检测卷积神经网络算法在 MIT-BIH 数据集上识别准确率为 94.7%，加速器电路的识别准确率为 89.7%。完成设计指标要求。

(3) 完成加速器电路设计，加速器电路部分功耗仅为 0.651W。完成设计指标要求。

表 5.4 实验结果与设计指标对比

参数	设计指标要求	实验测试结果
模型压缩率	50%	25%
算法识别准确率	90%	94.7%
加速器识别准确率	85%	89.7%
加速器电路功耗	1W	0.651W

5.5 本章小结

为了正确评估本课题所设计的基于卷积神经网络的新型生理信号检测算法硬件加速器电路，本章节进行了功能和性能的测试。在功能仿真中，对各个功能模块进行独立仿真和联合仿真，分析各个功能模块是否与第四章一致。实验结果表明，各模块均运行正常。在性能评估中，加速器电路最终功耗为 0.651W，单次识别运行时间为 4.4μs，硬件加速器识别准确率达到 89.7%。

第六章 总结与展望

6.1 总结

近几年来，神经网络算法在图像分类、语音识别等领域有着突出的表现。随着神经网络算法的不断发展，在医疗健康领域也成为了一个研究热点。针对生理信号的智能检测系统在医生诊断和个人健康监测中，开始扮演越来越重要的作用。医生可以根据智能诊断系统做出更准确的判断，患者可以通过便携式的生理信号检测系统来实时保障自己的身体健康和生命安全。

本文以基于卷积神经网络的新型生理信号检测算法及硬件加速器电路为研究课题，通过软硬件协同的方式，提出了一种高准确率、高压缩率、参数数量少的生理信号检测算法，并设计了对应的低功耗硬件加速器电路，主要工作包括：

(1) 以卷积神经网络与传统生理信号检测算法为理论基础，提出了一种新型生理信号检测算法，其中包含卷积层、池化层、激活函数和全连接层；使用全局平均池化层来代替一个全连接层，减少全连接层参数数量，和运算次数；

(2) 对提出的新型生理信号检测算法进行优化。主要采用参数量化的方式对模型进行压缩，由原先的 32 位浮点数量化为 8 位的定点数；并对量化结果进行了分析比较，实验结果表明，在识别准确率几乎没有下降的情况下，模型大小变为原先的 25%；

(3) 设计了新型生理信号检测算法所对应的硬件加速器电路。对脉动阵列、权重读写模块、数据读写模块的设计原理与计算原理进行了详细的阐述；优化了数据读写模块的数据流，降低了访存次数；

(4) 对设计的硬件加速器电路进行了功能仿真和性能评估。以功能划分，进行了模块内和模块间的独立仿真测试与联合仿真测试，最终加速器电路功耗仅有 0.651W；并针对 MIT-BIH 心律失常数据集进行了多平台准确率和运行速度测试，实验结果表明加速器识别准确率为 89.7%，单次识别运行速度为 4μs。

6.2 展望

本课题提出了一种基于卷积神经网络的新型生理信号检测算法，并设计了对应的硬件加速器电路，达到了设计指标的要求，但本课题仅完成了硬件加速器电路的设计，仍存在一些有待完善和改进之处。

首先，生理信号检测算法的性能仍有待提高。识别准确率方面，目前在 MIT-BIH 数据集 5 分类问题上最佳的心律失常分类算法识别准确率已经接近 100%，但是在 17 分类问题上仍有较大的提升空间。模型结构方面同样可以改进，卷积神经网络的特点决定了高参数量难以避免，而其他神经网络结构，诸如 LSTM 等，在生理信号检测方面表现同样优异，在后续工作中，可以结合多种神经网络的结构特点，减小网络深度，减少参数数量，设计混合型的生理信号检测算法。

其次，在量化过程中，模型大小仍可以进一步压缩。不同网络层对于量化的敏感程度是不同的，本课题将各层参数统一量化为 8bits，但是部分层仍可以继续压缩至 6bits，乃至 4bits，可以进一步实现更高的模型压缩率。

最后，本课题中，只设计了硬件加速器电路，并未设计完整的基于 FPGA 的生理信号检测系统，仅仅对电路完成了仿真和性能评估，对于系统完整实现和实际识别准确率、功耗等实验仍有待完善。这部分内容将在我研究生生涯中，进行更深入的学习与研究。

参考文献

[1] 胡盛寿, 高润霖, 刘力生, et al. 《中国心血管病报告 2018》概要 [J]. 中国循环杂志, 2019, 34(03) : 209-20.

[2] CAMM A J, KIRCHHOF P, LIP G Y H, et al. Guidelines for the management of atrial fibrillation The Task Force for the Management of Atrial Fibrillation of the European Society of Cardiology (ESC) [J]. Europace, 2010, 12(10) : 1360-420.

[3] BIYKEM B, E. H R, JAVED B, et al. 2021 ACC/AHA Key Data Elements and Definitions for Heart Failure: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Data Standards (Writing Committee to Develop Clinical Data Standards for Heart Failure) [J]. Circulation: Cardiovascular Quality and Outcomes, 2021, 14(4).

[4] HONG S, ZHOU Y, WU M, et al. Combining deep neural networks and engineered features for cardiac arrhythmia detection from ECG recordings [J]. Physiological Measurement, 2019, 40(5).

[5] PŁAWIAK P. Novel methodology of cardiac health recognition based on ECG signals and evolutionary-neural system [J]. Expert Systems With Applications, 2018, 92.

[6] WENG Y, ZHOU T, LIU L, et al. Automatic Convolutional Neural Architecture Search for Image Classification Under Different Scenes [J]. IEEE Access, 2019, 7: 38495-506.

[7] YıLDıRıM Ö, PŁAWIAK P, TAN R-S, et al. Arrhythmia detection using deep convolutional neural network with long duration ECG signals [J]. Computers in Biology and Medicine, 2018,32(3).

[8] PHILIP D C, MARIA O D, B R R. Automatic classification of heartbeats using ECG morphology and heartbeat interval features [J]. IEEE transactions on bio-medical engineering, 2004, 51(7).

[9] PARK K S, CHO B H, LEE D H, et al. Hierarchical support vector machine based heartbeat classification using higher order statistics and hermite basis function; proceedings of the 2008 Computers in Cardiology, F 14-17 Sept. 2008, 2008 [C].

[10] 侯晓晴, 仝泽友, 刘晓文. 基于改进小波变换的 QRS 特征提取算法研究 [J]. 现代电子技术, 2020, 43(13) : 57-61.

[11] MATHEWS S M, KAMBHAMETTU C, BARNER K E. A novel application of deep learning for single-lead ECG classification [J]. Computers in Biology and Medicine, 2018, 99.

[12] LIH O S, K N E Y, SAN T R, et al. Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats [J]. Computers in biology and medicine, 2018, 102.

[13] MENEGUITTI D F, L.M. M H, WULFERT C T, et al. Arrhythmia classification from single-lead ECG signals using the inter-patient paradigm [J]. Computer Methods and Programs in Biomedicine, 2021, 202.

[14] YANG W, SI Y, WANG D, et al. Automatic recognition of arrhythmia based on principal component analysis network and linear support vector machine [J]. Computers in Biology and Medicine, 2018, 101.

[15] VENIERIS S I, KOURIS A, BOUGANIS C-S. Toolflows for Mapping Convolutional Neural Networks on FPGAs: A Survey and Future Directions [J]. Acm Computing Surveys, 2018, 51(3).

[16] CHEN Z, LUO J, LIN K, et al. An Energy-Efficient ECG Processor With Weak-Strong Hybrid Classifier for Arrhythmia Detection [J]. IEEE Transactions on Circuits and Systems II: Express Briefs, 2018, 65(7) : 948-52.

[17] 李飞腾. 基于卷积神经网络的心电分类算法及高能效加速器架构研究 [D]; 浙江大学, 2020.

[18] 杨淑莹, 桂彬彬, 陈胜勇. 基于小波分解和 1D-GoogLeNet 的心律失常检测 [J]. 电子与信息学报, : 1-10.

[19] 急性 ST 段抬高型心肌梗死诊断和治疗指南 [J]. 中华心血管病杂志, 2015, 43(05) : 380-93.

[20] 高润霖. 急性心肌梗死诊断和治疗指南 [J]. 中华心血管病杂志, 2001, (12) : 9-24.

[21] 慢性稳定性心绞痛诊断与治疗指南 [J]. 中华心血管病杂志, 2007, 35(03) : 195-206.

[22] INAN O T, GIOVANGRANDI L, KOVACS G T A. Robust Neural-Network-Based Classification of Premature Ventricular Contractions Using Wavelet Transform and Timing Interval Features [J]. IEEE Transactions on Biomedical Engineering, 2006, 53(12) : 2507-15.

[23] LLAMEDO M, MARTíNEZ J P. Heartbeat Classification Using Feature Selection Driven by Database Generalization Criteria [J]. IEEE Transactions on Biomedical Engineering, 2011, 58(3) : 616-25.

[24] NASIRI J A, NAGHIBZADEH M, YAZDI H S, et al. ECG Arrhythmia Classification with Support Vector Machines and Genetic Algorithm; proceedings of the 2009 Third UKSim European Symposium on Computer Modeling and Simulation, F 25-27 Nov. 2009, 2009 [C].

[25] GRAVES A, MOHAMED A-R, HINTON G, et al. SPEECH RECOGNITION WITH DEEP RECURRENT NEURAL NETWORKS [M]. 2013 Ieee International Conference on Acoustics, Speech and Signal Processing. 2013: 6645-9.

[26] ZHENG S, JAYASUMANA S, ROMERA-PAREDES B, et al. Conditional Random Fields as Recurrent Neural Networks [M]. 2015 Ieee International Conference on Computer Vision. 2015: 1529-37.

[27] 孔祥溢, 王任直. 人工智能及在医疗领域的应用 [J]. 医学信息学杂志, 2016, 37(11) : 2-5.

[28] 张菊英, 韦健, 杨树勤. 神经网络模型在住院费用影响因素分析中的应用 [J]. 中华医院管理杂志,

2002, (03) : 18-20.

[29] WEI Y, ZHOU J, WANG Y, et al. A Review of Algorithm & Hardware Design for AI-Based Biomedical Applications [J]. IEEE Transactions on Biomedical Circuits and Systems, 2020, 14(2) : 145-63.

[30] ACHARYA U R, FUJITA H, LIH O S, et al. Automated detection of coronary artery disease using different durations of ECG segments with convolutional neural network [J]. Knowledge-Based Systems, 2017, 132.

[31] ARIF M, MALAGORE I A, AFSAR F A, et al. Automatic Detection and Localization of Myocardial Infarction using Back Propagation Neural Networks [M]. 2010 4th International Conference on Bioinformatics and Biomedical Engineering. New York; Ieee. 2010.

[32] UBEYLI E D. Recurrent neural networks employing Lyapunov exponents for analysis of ECG signals [J]. Expert Systems with Applications, 2010, 37(2) : 1192-9.

[33] LIN M, CHEN Q, YAN S. Network In Network [J/OL] 2013, arXiv:1312.4400 [https://ui.adsabs.harvard.edu/abs/2013arXiv1312.4400L.

[34] JACOB B, KLIGYS S, CHEN B, et al. Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference; proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, F 18-23 June 2018, 2018 [C].

[35] CHEN Y, EMER J, SZE V. Eyeriss: A Spatial Architecture for Energy-Efficient Dataflow for Convolutional Neural Networks; proceedings of the 2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA), F 18-22 June 2016, 2016 [C].

[36] MOONS B, GOETSCHALCKX K, BERCKELAER N V, et al. Minimum energy quantized neural networks; proceedings of the 2017 51st Asilomar Conference on Signals, Systems, and Computers, F 29 Oct.-1 Nov. 2017, 2017 [C].

致　　谢

时光飞逝，白驹过隙，大学本科四年即将过去，回顾在东南大学度过的本科四年时间，自己不论在专业知识还是工作能力方面都有很多收获。在这里，我要感谢所有陪我走过的大学四年的人，谢谢你们对我一直以来的帮助和支持。

首先，我要感谢我的父母和家人。父母在我在外求学时，不断给我支持，在我心情低落时给予我鼓励和经济上的支持。每次和妹妹聊天都会带给我温暖和快乐，他们永远是我最温馨的港湾。

然后，我要感谢我的导师刘昊。刘老师为人和善可亲，在专业知识方面也给了我许多指导，在开题、中期、实验和论文撰写的过程中都给了不少指导和修改意见，在我陷入歧路时及时指正，真正引领我进入了科研的大门。

同时，我也要感谢我的学长学姐们，感谢李支青、黄俊光和苏峰学长在我刚刚接触课题，内心迷茫时及时给予帮助，在实验过程中，他们的指导使我受益匪浅。我还要特别感谢我的舍友和朋友们，感谢孙礼、王腾熠、甄涵文和朱近赤在做毕设过程中互帮互助，感谢孔令行、闵成、徐律衡和徐智韩使我在南京度过了快乐的大学四年，也感谢傅雨凝、杨皓翔和魏宇恒等朋友的陪伴，谢谢你们。

最后，感谢各位评审老师，感谢你们参与我的答辩并评审我的论文，为我的大学四年画上完整的句号。也希望我在将来的研究生生涯中继续进步，始终健康、快乐、幸福。