基于 FPGA 和卷积神经网络量化的语音分类加速器*

温 冬，姜晶菲，窦 勇，许金伟，肖 滔
（国防科技大学计算机学院， 湖南 长沙 410073）

摘 要：传统的卷积神经网络在执行语音分类任务时存在模型存储规模大、浮点数值格式算力需求高的问题，限制了算法在低功耗和高吞吐率场景下的应用。通过对网络参数进行二值化处理、网络激活值进行定点数量化，降低了神经网络的存储体积并将浮点数运算转化为较快的定点数运算。基于该量化优化，进行软硬件协同设计，并利用 FPGA 平台开发出全流水、低功耗和高吞吐率的语音分类加速器，与先进多核 CPU 平台相比，单 PE 加速器在吞吐率上取得了 18-300 倍的加速比。

关键词：可重构计算; 神经网络; 量化压缩; 语音分类

中图分类号：TP391.9 文献标志码：A

A Speech Classification Accelerator Based on FPGA and Convolution Neural Network Quantization

WEN Dong, JIANG Jing-fei, DOU Yong, XU Jin-wei, XIAO Tao, HAN Zhe
(School of Computer Science, National University of Defense Technology, Changsha 410073, China)

Abstract: Traditional convolution neural networks have defect of huge parameter storage space and demand for extensive computing power due to floating data type when handling speech classification tasks, thus limiting it's use under low-power and high-throughput circumstance. Turning parameter into binary format and turning activation data into fix-point format can reduce neural network's storage space and replace floating computing with faster fix-point computing. Based on this optimization and software-hardware-cooperating work, we design a full-pipeline, low-power and high-throughput speech classification accelerator on FPGA platform, which has 18-300x throughput accelerating ratio with single PE compared to state-of-art multi-core CPU platform.

Keywords: programmable computing; neural network; quantization; speech classification

1 引言

收稿日期：****－**－**;；修回日期：****－**－**
* **基金项目**：核高基国家重大专项（编号：2018ZX01028101）
通迅地址：410073 湖南省长沙市国防科技大学计算机学院 姜晶菲 jingfeijiang@nudt.edu.cn
Address: School of Computer, National University of Defense Technology, Changsha 410073, Hunan , P.R.China

自从 Deng、Yu 等人[1]将 RNN（循环神经网络）和 LSTM（长短期记忆）模型引入声学问题领域以来，RNN 在语音识别和声音分类领域取得了一系列性能上的突破。然而，基于 RNN 的深度神经网络结构复杂，其大量循环时序计算难以训练和并行化，限制了 RNN 模型在实时场景下的应用。此外，CNN（卷积神经网络）在音频对象上具有同样优异的性能[2]。许多模型利用 MFCC（梅尔频率倒谱系数）或其它滤波算法从音频中提取音频特征，进而把音频帧和音频文件转化为特征图[3]，然后像计算机视觉任务中的卷积神经网络一样，设置卷积层、池化层、批量归一化层和全连接层，让卷积神经网络模型运行在音频特征图上。通过利用 3x3 这样的小尺寸卷积核，声音卷积神经网络相比长短期记忆深度神经网络可以更快地训练和推理，并且拥有更多的并行加速设计空间。

然而，传统高性能平台在部署卷积神经网络算法模型时存在两个重要的障碍: 功耗和速度。高性能 CPU 可以在比 GPU 更低的功耗下运行，但其较低的计算并行度无法高效地处理计算密集型的卷积运算。另一方面，高并行度的 GPU 可以更快地运行卷积神经网络，但它们高达数百瓦的功耗往往会对部署场景提出更多的限制。为了保证计算精度，深度神经网络中通常在运算和存储时通常采用浮点数据格式，但浮点格式在存储和计算这两个方面上都对 CPU 和 GPU 平台很不友好。浮点数据需要更大的存储字长和更多的计算功能部件，这就导致了更大的存储空间开销和更大的电路设计面积，进而增加了硬件功耗。同时，浮点数据格式的计算复杂度决定了设计者很难通过硬件设计或算法改进减少计算周期数。

一些工作[4]已经证明卷积神经网络在进行推理时并不一定需要浮点数据格式，低精度计算也可以让算法获得与全精度浮点数据相似的性能。这些工作为硬件设计人员提供了一些新的思路: 通过量化，研究人员在精度损失很小的情况下可以将权重和激活值数据转化为半精度浮点格式、8 位浮点格式、定点格式甚至二值化格式[5]。基于各种卷积神经网络的量化方法，学术界和工业界出现了 BNN（二值化神经网络）加速器和支持 8bit 浮点数据的 GPU 等。这些设计与传统的处理器平台相比，在大大降低了功耗的同时具有高达数百倍的加速比。基于卷积神经网络的声学算法和对硬件友好的卷积神经网络模型量化方法非常适合我们设计一个实时声音分类应用加速器。

与 GPU 和 CPU 平台相比，ASIC（专用集成电路）和 FPGA（现场可编程逻辑门阵列）高度定制化的设计特性使得开发者可以更灵活和更针对性地利用量化的卷积神经网络模型。ASIC 和 FPGA 通过设置与量化模型相同的数据位宽与数据格式、设计不同层次的流水线、扩展并行度，进而达到更好的计算性能和更低的运行功耗。在运行卷积神经网络模型时，ASIC 可以在计算性能和功耗上都取得很好的效果，但昂贵的设计和制造成本限制了它的广泛应用。而 FPGA 由于其可配置的特点和成熟的工业设计，可以在性能、功耗和成本之间保持较好的平衡，因此很适合作为卷积神经网络的加速平台。

在前人工智能时代，FPGA 被广泛用于数据中心的网络通信加速、云服务器操作系统虚拟化和生物科学计算加速等。而在人工智能时代，功耗低面积小的 FPGA 不仅可以为边缘端带来更多算力，其配置灵活的特性可使用户在云端实现人工智能算法进行定制化加速、高吞吐率负载和低延迟优化，为高性能计算拓展了新的应用场景。

声音分类是一类典型的情报分析任务之一，对特定集合中语音指令、战场命令等的快速识别是智能场景分析中的基础一环。目前利用典型的深度卷积神经网络模型可以较好地完成有限声音集合的粗分类，但模型还有利用量化技术进行压缩和优化的空间。为了进一步压缩声音分类应用的模型规模，更加充分地利用和发挥定制计算相较于传统高性能计算平台的优势，需要研究更加精简的定制化模型加速方法。基于此，我们在 Tensorflow 语音数据集中选取构造了一种典型的声音命令集合，并基于一种典型的声音分类卷积神经网络模型进行定制优化压缩设计。该网络权值为+1 或-1，原始的激活数据为全精度浮点数据格式，网络由两个卷积层和三个全连接层构成[6]。我们用 FPGA 硬件资源实现了该二值权重网络（BWN）声音分类模型的定制加速器，与目前先进的 CPU 平台相比，实现了 18-300 倍的吞吐率加速比和较低的运行功耗。本文的主要工作如下:

我们利用 Matlab 的内置量化函数和设计空间探索方法将一个声音分类卷积神经网络的浮点型激活值数据逐层地转化为定点数据，FPGA 平台上的定点计算性能弥补了定点数据的精度损失。

我们设计了一个基于 FPGA 的多 PE BWN 加速器，该加速器具有共享权值存储设计、均

衡流水线结构和卷积神经网络层间低延迟流水线设计。

我们在 CPU 平台上进行了单线程、多线程和多节点环境下的目标声音分类模型的性能测试，得到了多组高性能 CPU 平台上的基准性能数据。与上述测试结果相比，我们的设计在性能功耗比和计算加速比上具有绝对优势。

2 神经网络推理量化

现在的深度神经网络模型通常包含大量的参数，为了更好地训练神经网络模型，研究人员通常选择全精度浮点数据格式作为参数数据格式。然而，在推理任务中，所有这些参数一旦训练完成在数值上就不会被改变，因此就不需要继续保留全精度数据格式或保留所有的参数。通过离线的对神经网络参数进行剪枝、量化压缩精度，我们可以显著的降低网络的存储体积、总计算量和计算延迟。[7] 提出了一种将浮点深度神经网络参数压缩为包含+1 和-1 二进制数据的算法。与原有的 64 位浮点数据格式模型相比，这种压缩方法不仅大大减少了参数存储空间需求，而且用加减运算替代卷积中的乘法运算，显著降低了单个卷积运算的计算延迟。该方法使设计人员可以将所有参数都放在芯片内，降低了内存数据交换开销；另外对于 FPGA 硬件平台而言，加减电路也比乘法器更易设计和实现。

另外，[7]中提出了将神经网络运行中的激活值数据也全部转化为二值格式的方法。与训练完成后不发生改变的神经网络参数不同的是，网络激活值数据会随着输入对象（如图片或音频特征图）的不同而产生数值上的波动，对激活值数据做二值化处理就会产生较大的精度损失。此时对网络激活值进行定点化处理就可以在精度和计算性能上得到较好的平衡：一方面定点数计算的计算周期数较浮点数大幅减少；另一方面定点格式可以根据激活值数据实际的数值分布范围来调整整数和小数位宽，小数位宽越大，对原数据的精度保持就更好，整数部分位宽则决定了数据格式的数值表征范围。相比二值化数据格式，定点数据可以更好的在计算精度和计算速度上达到平衡。

3 声音分类模型

3.1 模型结构和权值二值化

该卷积神经网络声音分类模型是在 Tensorflow 语音指令数据集上进行训练得到，能够分辨出“up”，“down”，“yes”，“right”，“left” 共 5 种单词语音段和未知声音(unknown)。最初得到的原始网络模型中的权值和中间结果（激活值）都由浮点格式表达。首先，该模型利用 MFCC 算法将一个音频文件变换成维度为 20*49*1 的浮点格式的张量，该张量即为音频的特征图矩阵。此张量将被送入包含两个卷积层、三个全连接层的卷积神经网络中，所有的卷积核大小均为 3，卷积步长为 1。该卷积神经网络没有边缘扩展，便于我们加速，该网络的结构如图 1 所示。最后，该模型通过 softmax 函数输出六种类型标签的预测概率。

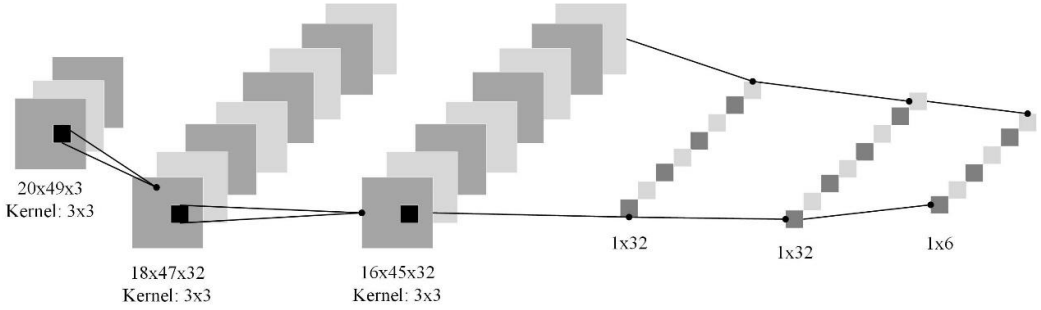

Figure 1 Convolution Neural Network's Architecture

图 1 卷积神经网络结构

当模型训练完成、参数数值都确定之后，我们利用 tanh 函数将数值分布没有约束范围的浮点权值约束到（-1，1）的范围，然后使用一系列放缩与映射的方法将浮

点权值离散为 0 或 1，最后离散为-1 或+1。 图 2 详细展示了我们处理权值数据的流程。经过测试，此时的模型（权值为二进制-1、+1，激活值为浮点数）的整体精度不低于 85%。

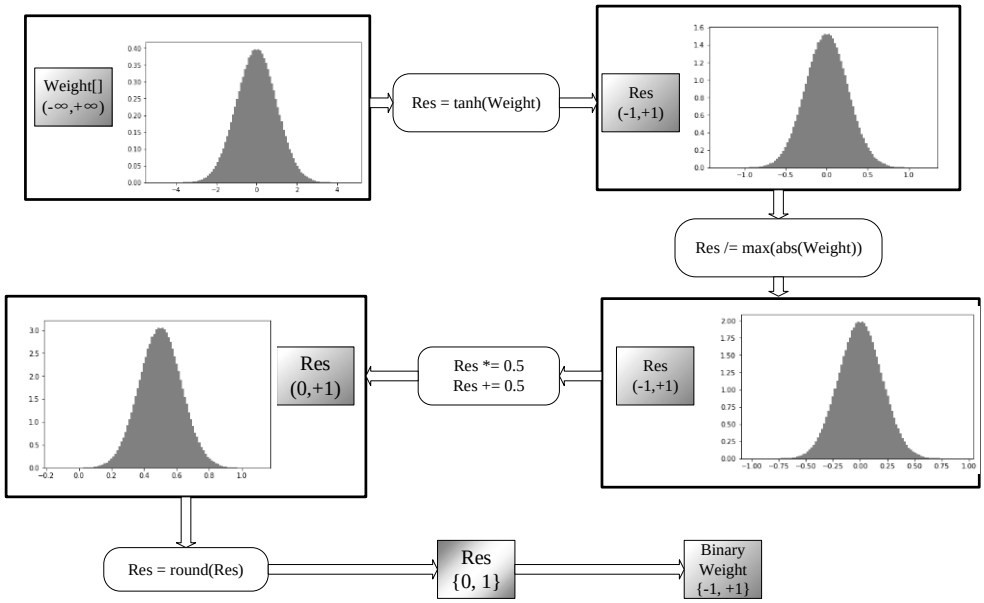

Figure 2 The Process of Parameter's Binarization

图 2 权值二值化流程

3.2 激活值数据量化

神经网络中每一层计算的数据可分为两部分：乘累加运算后的中间结果、中间结果进行批量归一化后的最终输出数据。通常来说，中间结果的绝对值和方差可能很大，但经过批量归一化处理后，方差和数据分布范围都将缩小，最终输出数据的分布情况会得到改善。[8]论证了批量归一化对于模型总体精度的重要影响，而批归一化操作中的乘除法不仅依赖先验的平均值和方差参数，而且对于数值精度较为敏感，因此为了兼顾硬件性能与整体精度，我们需要对同一层内的中间结果和输出结果采取不同的量化数据位宽以尽可能保证批归一化的计算精度。

为了将激活数据转化为定点格式并使精度损失尽可能小，我们首先确定每个层内中间结果和输出结果的数值分布范围（即数据整数部分位宽），然后利用Matlab量化函数库和设计空间搜索法确定每个层中两种数据的最佳小数位宽量化格式组合。所以，我们提出了一种基于权值二值化、激活数据定点化的卷积神经网络声音分类模型。在本文的实验部分将讨论具体的数据量化结果。

4 加速器体系结构

由于目标加速网络体积较小、层数较浅且权值都是二值化的，所以我们主要设计重点是设置共享的片上二值化参数存储器、逐层加速神经网络以及在层间设计均衡流水线提高加速器的整体性能。

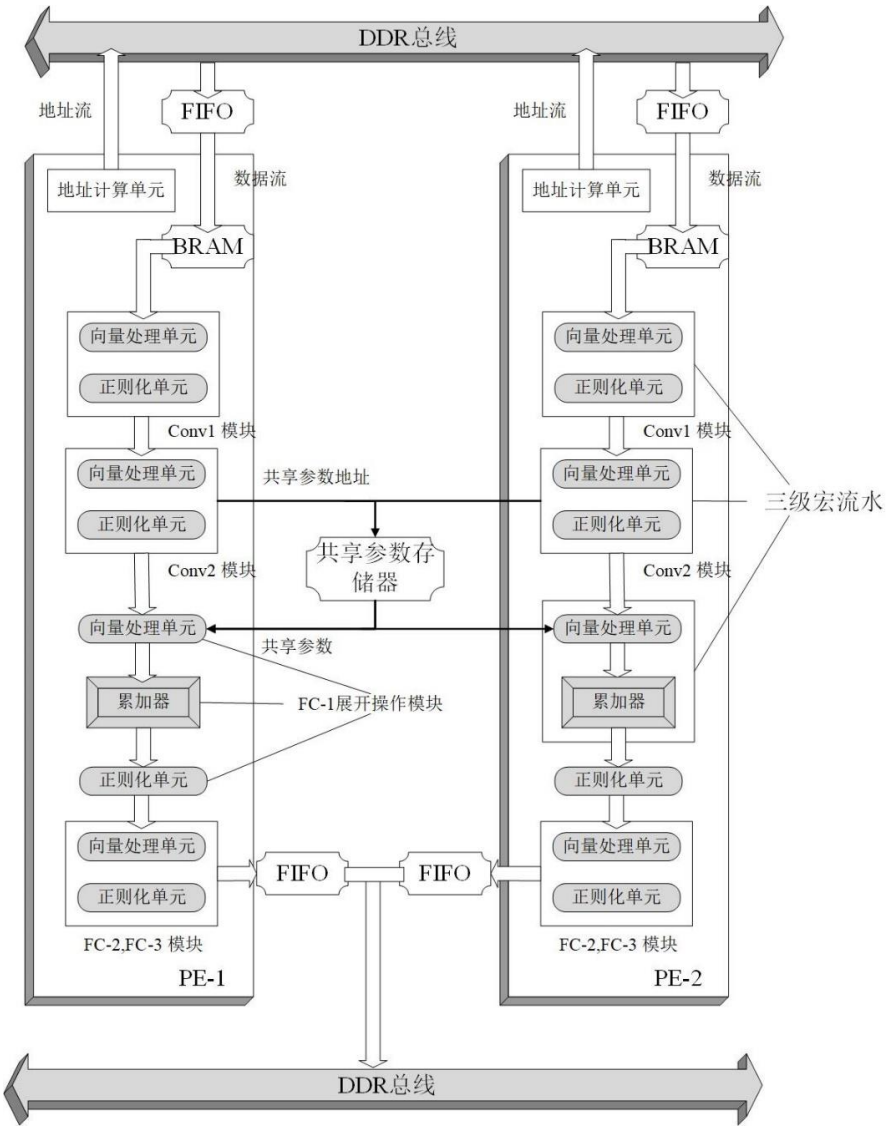

Figure 3 Hardware Architecture

图 3 硬件结构图

一个加速器可以包含一个或多个 PE、一个 DDR 总线和一个 DRAM 存储器。 每一个 PE 都有自己的功能模块，包括五层神经网络处理单元、地址计算模块、输入输出与数据存储。其中卷积层 1 模块、卷积层 2 模块和全连接层 1 模块按照层间流水线的方式工作。

4.1 参数存储

与一些加速较大规模神经网络的加速器需要用 DRAM 辅助片上存储器来存储参数不同，我们的模型参数规模适中，在采用了二值化处理后参数存储空间进一步减少。 表 1 显示了该模型参数的详细信息。

Table 1 Parameter Information

表 1 参数信息表

层数	滤波器数量	卷积核大小	参数量	参数存储大小
conv1	32	3*3*1	288	36B
conv2	32	3*3*32	9216	1152B
fc1	32	16*45*32	737280	92160B

fc2	—	—	1024	128B
fc3	—	—	192	24B

全连接层 1 占据了绝大部分的网络参数，其他层的参数数据规模相当小，可以直接存储在片上。考虑到全连接层 1 参数的巨大规模，一个很自然的想法是在多个 PE 之间共享它们。我们设置所有 PE 为同步工作，并在全连接层 1 的计算时提取完全相同的预训练参数。该共享存储模块由 32 个 BRAM 块组成，每个 BRAM 块用于存储全连接层 1 中的一个滤波器参数。

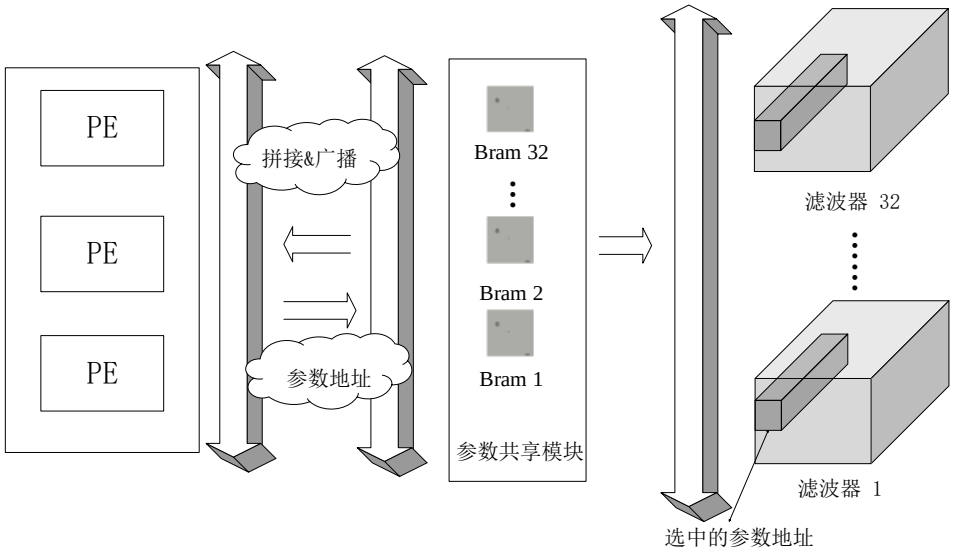

Figure 4 FC-1's Sharing-Parameter Design

图 4 全连接层 1 的参数共享结构

PE 向参数共享模块发送目标参数地址，该地址是 BRAM 的行地址。BRAM 中的参数沿第三维存储，其存储方式与全连接层 1 的输入数据组织方式一致。 通过广播，共享参数数据被发送到所有 PE。

对于除了全连接层-1 之外的二值化网络参数，我们将其直接存储于片内，每个 BRAM 块对应存储一个滤波器中的所有参数。每次运算时，BRAM 块都将对应像素位置的参数直接交付向量处理单元并与输入激活值进行向量运算。向量运算器首先根据传入的 1bit 参数（代表+1，-1）对激活值进行原样输出或是取反操作得到 32 个参数运算结果，再利用加法树对 32 个激活值对应的参数运算结果进行逐层累加，最终得到该次运算的结果。

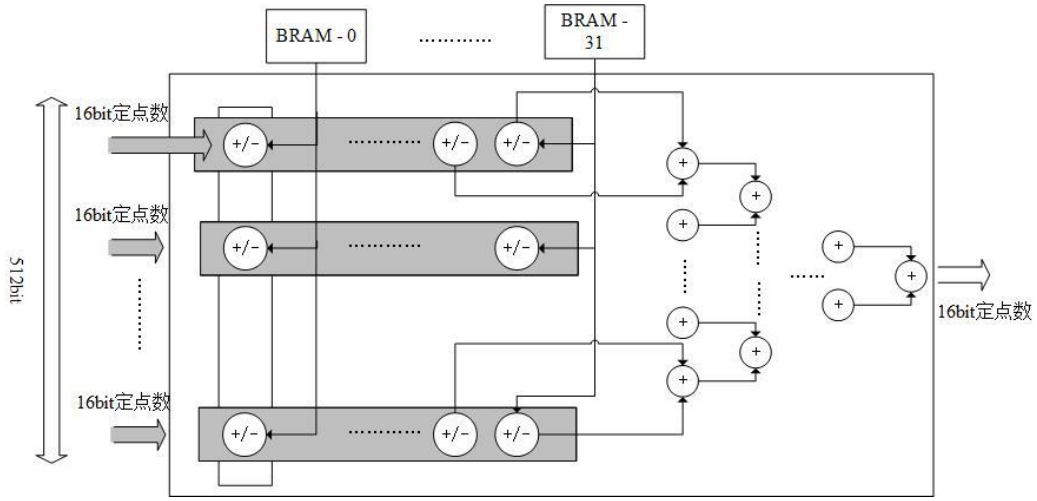

Figure 5 Vector Processing Unit

图 5 向量处理单元

4.2 位宽扩展

在一个卷积层或全连通层内，输入的定点激活数据往往是经过了批归一化的，因此通常服从正态分布、数值分布范围较小，而经过乘累加计算后，数据的方差可能会很大且不规则。批量归一化依赖于这些中间结果的均值和方差以改善乘累加后的数据分布，这对最终的精度至关重要。实验结果证明，如果在归一化步骤中方差和均值数据使用与输入激活值相同的量化数据格式，将会造成一定的数值精度损失并给模型的分类准确率带来不必要的损失。

为了解决这一问题，我们引入位宽扩展以改善计算精度：对于批归一化操作中的平均值、方差超参数，我们增加额外的小数位宽以增加此类参数在硬件上的表示精度；对于精度需求相对不高、小数位宽相对较少的计算中间结果，我们在批归一化操作时对其做简单的小数位宽扩充，使其能够和平均值、方差参数对齐。当 DSP（数字信号处理器）输出归一化运算结果后，加速器再对多出的小数位宽进行截断，使数据恢复到原始的输入激活值量化格式。

同样的，乘累加运算的中间结果也有类似的位宽扩展需求。累加器对多组乘法运算的乘积进行累加的过程中可能会出现数值溢出的情况，因此乘累加的结果需要更多的整数位宽而不是小数位宽。我们的加速器对归一化步骤和乘累加结果累加器进行了扩展，实验结果表明该方法较好的保证了计算精度。

4.3 层间流水线

实现层间流水线的关键是要平衡流水线每一层间的运行周期数。在 VGG-16 和 AlexNet 等深度卷积神经网络中通常很难保持这种平衡，因为随着网络的深入，更深的层需要前面的层以更快的速度生成激活值结果，这一设计要求大大超出了当前硬件平台的计算能力。

本文的目标加速网络层数较少、深度较浅，所以相比 VGG 和 Alex 网络更容易在两个卷积层之间保持平衡。我们改变了卷积层 1 的取数据地址生成方法，使之与卷积层 2 的计算模式相匹配。同时，为了能在运算周期数上实现两个卷积层的均衡，我们扩展了卷积层 1 功能模块的并行度，使加速器可以在一个流水线宏节拍内生成一组可被卷积层 2 完整使用的输入数据。

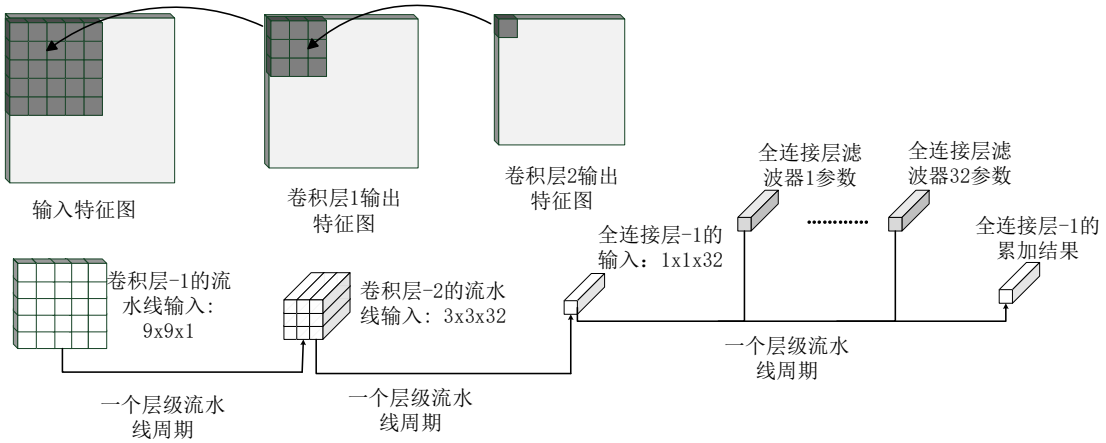

Figure 6 Balancing Pipeline Between Conv-1 and Conv-2

图 6 平衡卷积层 1 和卷积层 2 之间的层流水线

为了在卷积层 2 中输出一组数据，网络模型需要卷积层 1 在输入音频特征图一个的 5*5 区域中计算 9 个 3*3 滑动窗口。为了平衡计算周期数，我们扩展了卷积层 1 的计算阵列规模，使其在一个宏流水线周期内产生 9 个 32*1 向量，在下一个周期内

这些向量将被发送到卷积层 2，并为全连接层 1 产生一个 32*1 向量，同时卷积层 2 的计算功能也将在一个宏周期内运行完毕。全连接层 1 是该加速器的瓶颈，由于计算量巨大，要保持该层与两个卷积层的平衡所需要的计算资源是当前平台难以承受的，但图 4 中的设计可以保证全连接层 1 在每个宏周期内保持一个向量进行累加。

通过对层流水线的调整，目标神经网络可以在不需要将层间结果写回 DRAM 中的情况下进行加速，从而降低存储开销，即从输入音频特征到最终预测结果的数据流始终保持在层次流水线中，加速器只在取数和最后回写时与 DRAM 通信。在流水线的各个层中，我们将所有的计算分解为向量计算单元、归一化单元等功能部件，提高了硬件的运行频率。

5 实验结构及分析

5.1 量化模型性能

我们利用 Matlab-2018a 的量化函数将激活值数据和批量归一化参数以饱和溢出方式转换为定点数据。为了找到最佳的位宽设置方法，我们进行了多组量化实验，并比较了它们的精度性能。

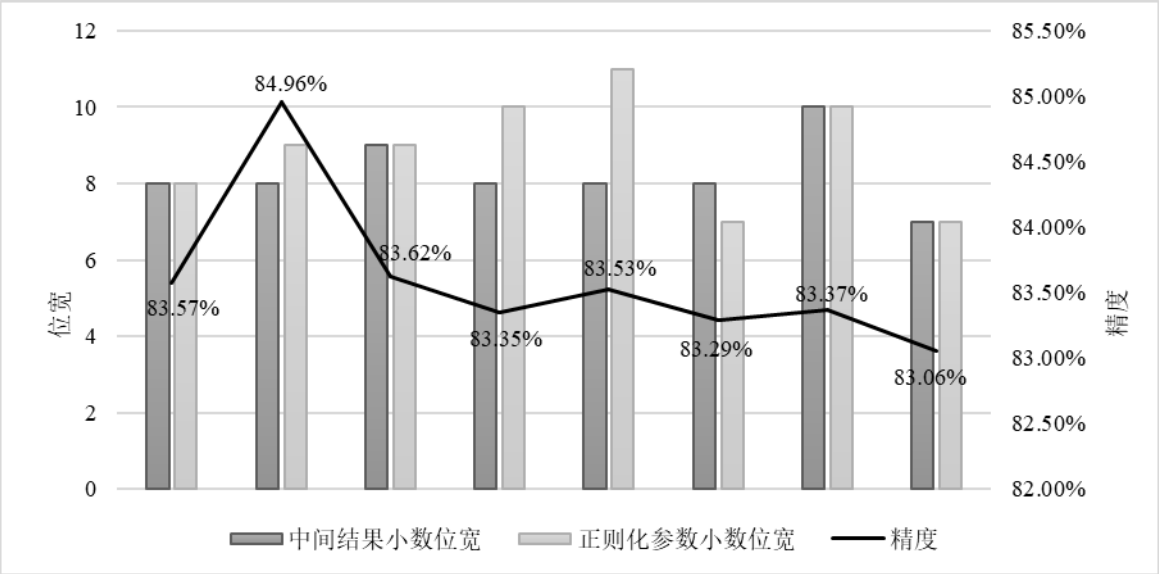

Figure 7 Accuracy Experiment Result

图 7 精度实验结果

考虑到数据格式设计要与硬件平台相契合，我们限定所有的数据格式位宽在 32bit~16bit 之间，并以此为限定条件对中间结果与归一化参数的位宽进行设计空间探索。我们发现，中间结果的整数部分位宽往往在 8~12bit 之间，而正则化参数中方差的整数位需要设置在 20~21bit 之间，由此确定了设计空间搜索中的小数位宽上界。在神经网络的量化算法中，激活值（中间结果）往往对数值精度有较高的鲁棒性，因此在寻找最佳方案时不需对中间结果的小数位宽做过多搜索。此外，考虑到归一化参数的精度需求比中间结果更高，因此在探索位宽组合时，一般设置归一化小数位宽大于等于中间结果小数位宽，但我们仍然在探索中做了与此设想相反的实验，实验结果验证了设想的正确性。

图 7 中的实验结果表明，当激活值数据采用 8 位小数位宽，归一化参数采用 9 位小数位宽时，该模型的性能最佳。需要注意的是硬件不能像 Matlab 代码那样简单地处理除法运算，因此我们将归一化函数中方差的除法转化为倒数乘法，并扩展小

数位宽以保证精度。

我们在 8700K 平台上对量化版本的神经网络程序进行分段计时。我们使用 matlab 单线程和多线程并行库运行程序，在不考虑 MFCC 预处理段的情况下，卷积层 2 是 CPU 平台上的主要性能瓶颈，对程序性能有着决定性的影响，而参数的量化操作则几乎对性能没有影响。

Table 2. Running Performance on CPU Platform

表 2. CPU 平台上的运行计时

8700K 单线程		
功能段	时间(Seconds)	时间占比
数据加载与预处理	9.413601	7.40%
卷积层 1 量化	0.000508	<1%
卷积层 1	3.084229	2.43%
卷积层 2 量化	0.000694	<1%
卷积层 2	114.582	90.14%
全连接层 1 量化	0.006863	<1%
全连接层 1	0.023377	<1%
全连接层 2 量化	0.000834	<1%
全连接层 2	0.001659	<1%
全连接层 3 量化	0.000945	<1%
全连接层 3	0.001589	<1%
总计	127.1163	100%
8700K 多线程并行		
功能段	时间(Seconds)	时间占比
数据加载与预处理	9.804942	19.74%
卷积层 1 量化	0.000195	<1%
卷积层 1	1.803836	3.63%
卷积层 2 量化	0.000307	<1%
卷积层 2	38.02353	76.56%
全连接层 1 量化	0.006184	<1%
全连接层 1	0.025257	<1%
全连接层 2 量化	0.000588	<1%
全连接层 2	0.001316	<1%
全连接层 3 量化	0	0
全连接层 3	0.000981	<1%
总计	49.66721	100%

5.2 硬件加速性能

我们在 8700K 和 Matlab-2018a 单线程平台、8700K 和 Matlab-2018a 并行库平台和 Intel Xeon 5220(2.2GHz,18 核)多节点 MPI 环境下（使用 Matlab-2018a 分布式并行库）运行整个数据集共 1512 个音频段。考虑到我们的加速器只负责加速卷积神经网络部分的运算，不涉及程序预处理以及音频帧转化为音频特征图的部分，因此在实验中我们只测试神经网络前向传播所消耗的时间，并不计入程序预处理与 MFCC 算法的部分。我们以上述实验环境的运行时间为基准来计算我们设计的加速比。表 4 展示了我们的测试结果，测试结果表明与当前先进的 CPU 平台相比，我们的加速器可以取得 18-300 倍的吞吐率加速比和更低的功耗。

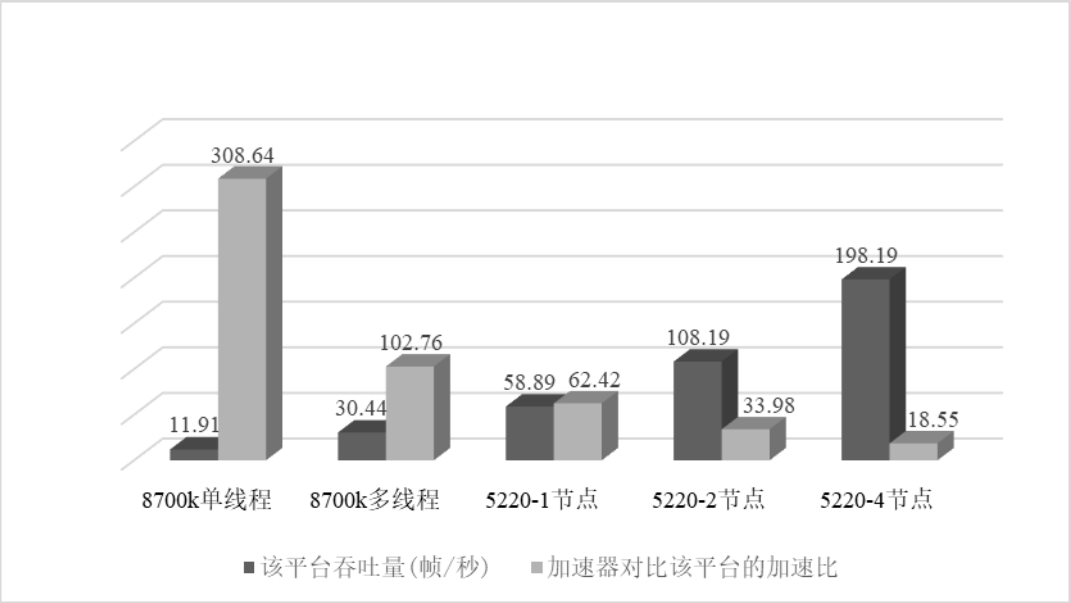

Figure 8 Performance Compared with Multiple Platforms

图 8 多平台性能对比

我们在 Xilinx KU-115 FPGA 和 Xilinx Vivado 2018.3 平台上实现了我们的单 PE 版本加速器设计，并测试了运行时间、运行精度和运行功耗。为了更好的测试上述数据，我们设置了加速器循环执行次数，进行了最长达 2 小时的压力测试。

Table 3 Accelerator's Performance

表 3 加速器性能参数

运行数据全集时间(*S*)	0.4116
静态功耗（W）	9.63
动态功耗（W）	10.06
运行准确率	84.96%
FPGA 平台资源利用率	38%
吞吐量(音频每秒)	3675.9
运行频率（MHz）	150
峰值定点运算次数(GOPS)	23.85

表 3 的数据表明卷积层 2 是最耗费时间的功能段，然而通过使用平衡的逐级流水线，我们的加速器可以消除原有神经网络中的瓶颈，获得优异的加速性能。 同时，流水线内的数据流大大减少了 DDR 总线的通信量，消除了其他计算平台上内存带宽限制。

Table 4 Comparison with Typical FPGA-Based RNN Accelerator

表 4 与典型的基于 FPGA 的 RNN 加速器的对比

加速器	RNN-LM[9]	RNN-Zynq[10]	BWN 加速器
硬件平台	FPGA	FPGA	FPGA
运行频率(MHz)	150	142	150
功耗(W)	25	1.942	10.06
峰值计算性能(GOPS)	9.6	0.2644	23.85
峰值吞吐量(FPS)	65.85	1073	3675.9
功耗比(GOPS/W)	0.38	0.1459	2.37

在表 4 中，我们选取两种典型的基于 FPGA 的 RNN 声学模型加速器并进行性能对比。我们的 BWN 加速器定制加速了卷积神经网络声学模型，在功耗比、吞吐率、峰值计算性能和功耗上都较 RNN 加速器有较大的优势。

6 结论

我们的加速器是一款针对特定声音分类模型的低功耗、高加速比、高吞吐量的专用硬件平台。我们的加速器在设计中突出了参数共享存储、精度灵活扩展和均衡的神经网络层间流水线的特点，通过软件算法与硬件平台的协同改进和设计使本加速器平台与现有的 CPU 平台相比具有优异的性能。我们在 Xilinx FPGA 上实现了设计方案并进行了压力测试，结果表明该加速器是一种可靠且高效的智能计算器件。

参考文献:

[1] Geoffrey H, Li D, Dong Y, et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition [J]. IEEE Signal Processing Magazine, 2012, 12:82-98.

[2] Tom S, Vaibhava G. Advances in Very Deep Convolutional Neural Networks for LVCSR[C/OL]. arXiv:1604.01792v2[cs.CL]

[3] Pakyurek M, Atmis M, Kulac S, et al. Extraction of Novel Features Based on Histograms of MFCCs Used in Emotion Classification from Generated Original Speech Dataset[J]. Electronics & Electrical Engineering, 2020, 26:46-51.

[4] Tung F, Mori G. Deep Neural Network Compression by In-Parallel Pruning-Quantization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(3):568-579.

[5] Lu C, Xu W, Jin S, et al. Bit-Level Optimized Neural Network for Multi-Antenna Channel Quantization[J]. IEEE Wireless Communications Letters, 2020, 9(1):87-90.

[6] Bo L, Hai Q, Yu G, et al. EERA-ASR: An Energy-Efficient Reconfigurable Architecture for Automatic Speech Recognition with Hybrid DNN and Approximate Computing[J]. IEEE ACCESS, 2018, 6:52227-52237.

[7] Matthieu C, Itay H, Daniel S, et al. Training Deep Neural Networks with Weights and Activations Constrained to +1 or -1[C/OL]. arXiv:1602.02830v3[cs.LG].

[8] Shibani S, Dimitris T, Andrew I, et al. How Does Batch Normalization Help Optimization[C/OL]. arXiv: 1805.11604v5[stat.ML].

[9] Sicheng L, Chunpeng W, Hai L et al. FPGA Acceleration of Recurrent Neural Network Based Language Model[C]. 2015 IEEE 23rd Annual International Symposium on Field-Programmable Custom Computing Mahcines, pp. 111-118.

[10] Andre C, Berin M, Eugenio C. Recurrent Neural Networks Hardware Implementation on FPGA[C/OL]. arXiv: 1511.05552v4[cs.NE]

作者简介：

稿件编号：

******

<table>
<tr><td>姓名</td><td>温冬</td><td>出生年</td><td>1997</td><td>性别</td><td>男</td><td>籍贯</td><td colspan="2">江苏省徐州市</td><td rowspan="3">  </td></tr>
<tr><td>学历</td><td colspan="8">硕士生 Master Candidate</td></tr>
<tr><td>职称</td><td>无</td><td>职务</td><td colspan="2">无</td><td>CCF 会员号</td><td colspan="3">无</td></tr>
<tr><td>研究方向</td><td colspan="9">可重构计算，计算机体系结构 Programmable Computing, Computer Architecture</td></tr>
<tr><td>联系电话</td><td colspan="5">18052192848</td><td>E-mail</td><td colspan="3">310-we-aaa-1@163.com</td></tr>
</table>

通讯地址	湖南省长沙市国防科技大学计算机学院 School of Computer, National University of Defense Technology, Changsha, Hunan	邮政编码	410073

姓名	姜晶菲	出生年	****	性别	女	籍贯	内蒙古自治区包头市	
学历	**博士 Ph.D**							
职称	研究员 Researcher	职务	教研室主任		CCF 会员号		无	
研究方向	计算机体系结构，人工智能芯片设计 Computer Architecture, Design of Artificial Intelligence Chip							
联系电话	18670778436				E-mail		jingfeijiang@nudt.edu.cn	
通讯地址	湖南省长沙市国防科技大学计算机学院 School of Computer, National University of Defense Technology, Changsha, Hunan					邮政编码		410073
姓名	窦勇	出生年	1966	性别	男	籍贯	吉林省舒兰市	
学历	**博士 Ph.D**							
职称	研究员 Researcher	职务	教研室主任		CCF 会员号		无	
研究方向	高性能计算，人工智能 High Performance Computing, Artificial Intelligence							
联系电话	13508476562				E-mail		yongdou@nudt.edu.cn	
通讯地址	湖南省长沙市国防科技大学计算机学院 School of Computer, National University of Defense Technology, Changsha, Hunan					邮政编码		410073
姓名	许金伟	出生年	1990	性别	男	籍贯	河南省周口市	
学历	**博士 Ph.D**							
职称	无	职务	无		CCF 会员号		无	
研究方向	可重构计算，人工智能 Programmable Computing, Artificial Intelligence							
联系电话	17873588397				E-mail		jinwei200911@163.com	
通讯地址	湖南省长沙市国防科技大学计算机学院 School of Computer, National University of Defense Technology, Changsha, Hunan					邮政编码		410073
姓名	肖滔	出生年	1997	性别	男	籍贯	湖南省衡阳市	
学历	**硕士 Master Candidate**							
职称	无	职务	无		CCF 会员号		无	
研究方向	可重构计算，计算机体系结构 Programmable Computing, Computer Architecture							
联系电话	15274914803				E-mail		xt@nudt.edu.cn	

通讯地址	湖南省长沙市国防科技大学计算机学院 School of Computer, National University of Defense Technology, Changsha, Hunan	邮政编码	410073