面向在轨高效实时图像处理的二值权重沙漏网络加速器设计

程利甫[1,2],徐炜莉[3],赵启义[1],段欣欣[2],蒋仁兴[2]
(1.清华大学 集成电路学院,北京 100047;2.上海航天电子技术研究所,上海 201109;
3.上海市宇航学会,上海 200235)

摘　要: 为了实现在轨高速实时图像处理,提出了基于多级预测校准机制及查找表方法的二值权重沙漏网络加速器。首先,提出了一种计算架构来统一支持二值权重及多位宽权重卷积计算;其次,提出了多级预测模型来实现可变精度的高效卷积,基于这2种方法可以将网络的计算量降低77.4%,推理速度提升2.3倍。本文针对残差模块中跳转连接造成的存储访问问题,提出了基于模块计算的流水架构,使得片上存储需求及访存操作分别降低了60%和31%,最终在28 nm工艺下完成了硬件后端设计以及性能分析,该加速器的面积为0.7 mm^2。在500 MHz工作频率下,功耗为117 mW,功耗效率达到10.15 TOPS/W,与当前主流二值加速器相比提升2个数量级以上。

关键词: 在轨图像处理; 沙漏网络; 查找表; 残差模块; 加速器
中图分类号: TN 332.1　　**文献标志码:** A　　**DOI:** 10.19328/j.cnki.2096-8655.2021.04.008

Binary-Weight Hourglass Network Accelerator for Efficient Real-Time Image Processing in Orbit

CHENG Lifu[1,2], XU Weili[3], ZHAO Qiyi[1], DUAN Xinxin[2], JIANG Renxing[2]
(1.Department of Microelectronics and Nano Electronics, Tsinghua University, Beijing 100047, China;
2.Shanghai Aerospace Electronic Technology Institute, Shanghai 201109, China;
3.Shanghai Society of Astronautics, Shanghai 200235, China)

Abstract: In order to achieve high-speed and energy-efficient processing for in-orbit images in real time, a binary-weight hourglass network (B-HN) accelerator based on multi-level prediction-correction mechanism and look-up table (LUT) approach is proposed. First, an LUT with a unified mode is adopted to support the convolutional neural networks with fully variable weight bit precision. Second, a multi-level prediction-correction model is proposed to achieve the computational-efficient convolution with adaptive precision. With these two methods, the computation amount of the network can be reduced by 77.4%, and the reasoning speed can be increased by 2.3 times. Third, a pipeline architecture based on block computing is designed for solving the memory access problem caused by the numerous skip connections in the residual block of B-HN. With the architecture, the on-chip memory requirements and access operations are reduced by about 60% and 31%, respectively. Finally, the hardware back-end design and performance analysis are completed under TSMC 28 nm technology. The area of the accelerator is 0.7 mm^2. At the operating frequency of 500 MHz, the power consumption is 117 mW, and the power efficiency is up to 10.15 TOPS/W, which are two orders of magnitude higher than any of current mainstream binary accelerators.

Key words: in-orbit image processing; hourglass network; look-up table; residual block; accelerator

收稿日期: 2020-09-19; **修回日期:** 2021-01-11
基金项目: 上海市优秀技术带头人计划(19XD1431400)
作者简介: 程利甫(1985—),男,博士研究生,高级工程师,主要研究方向为星载微系统、综合电子、有效载荷的研发应用。

0　引言

以卷积神经网络为代表的深度学习方法，已经在很多图像处理领域取得了超越人类的性能而获得广泛的应用，并随着航空航天领域图像传感技术的演进逐步渗透到该领域内，如在轨图像处理[1-3]、遥感目标识别[4-7]等。沙漏网络（Hourglass Network，HN）作为一种经典的网络架构被广泛应用于特征点检测等领域，但较高的计算复杂度和存储需求限制了其在设备端的应用。作为有效降低计算复杂度的一种方法，二值权重处理通过将权重值转换为＋1或－1，使得原来的乘法操作变为加法操作。

本文对沙漏网络的权重进行二值化处理得到二值权重沙漏网络（Binary-weight Hourglass Network，B-HN）。虽然可以将沙漏网络的权重存储需求降低到1/32，并且将计算量降低了近1/2，但是，推理过程中的算术操作数量仍然很高，需要对B-HN算法进行进一步优化。因此，本文的主要设计目标是在不影响识别精度的前提下，通过算法-硬件协同设计实现高能效的高速图像处理。为了避免造成比较大的精度损失，B-HN网络中第一层以及最后一层权重需要保持为多比特位宽（硬件实现时量化为8 bit）。

针对在计算和存储两方面潜在的瓶颈问题，采用多级预测校准机制及查找表（Look-Up Table，LUT）方法设计实现B-HN加速器来满足高效实时处理需求。

首先，采用LUT的查表方式来代替大量重复的加法操作。在UNPU[8]中，LUT被设置为两种模式：单比特模式和多比特模式来实现不同的计算效率。为了进一步提升硬件效率，本文将多比特权重进行离线编码，转换成可复用单比特权重的形式，从而用一种统一模式实现对不同位宽权重计算的支持，可以进一步降低硬件实现开销并且获得更高的能量效率。其次，作为B-HN中广泛采用的基本结构——“卷积-归一化-ReLU”，在经过激活后的激活值稀疏度会达到30%～90%。

卷积后为负值的结果在经过激活后会被直接置零。因此，针对这些结果为负值的卷积，如果能够提前判断其为负的话，则没必要执行全部完整精确的计算过程。同样，对于最大值池化而言，只需要在热点图中定位可能为最大值的点的位置即可。卷积过程中的部分计算结果就可以用来提前中止最终为负值的计算或提前滤掉非极大值的点。

基于这些发现，提出了多级预测校准机制来实现对计算精度的自动调整。与基于内核方法[9-10]相比，该方法在AlexNet和VGG-16网络上可以降低77.4%～82.8%的计算量，并分别取得2.3～3.4倍推理速度提升。

由于残差网络模块中存在大量的模块内及模块间的跳转连接，如果采用传统的逐层计算模式会造成大量的存储访问操作。而频繁的片外数据访问会造成巨大的功耗开销。为了解决这一问题，提出了基于模块计算的流水架构（Block Computing Based Pipeline，BCP）来提高片上数据重用。与传统的逐层计算模式[11]相比可以降低66.2%的片外数据访问。与相似的合并层方法[12]相比，分别降低了60%的片上存储需求和31%的数据访问。

在以上优化方法的基础上，在TSMC 28 nm CMOS工艺下对B-HN加速器实现了后端设计和性能分析。该加速器在500 MHz工作频率下的功耗为117 mW，功耗效率达到10.15 TOPS/W。

1　二值权重沙漏网络模型优化

1.1　沙漏网络模型简述

基于沙漏模型的算法在特征点检测的灵活性及精度方面具有明显的优势。如图1所示，沙漏网络在自底向上以及自上向下处理过程中呈现对称分布的特点，因此，可以获取图像在各个尺度的关键信息。

除此之外，还可以通过不断重复叠加沙漏模块来提高精度。本文针对沙漏网络进行二值权重优化，在通过B-HN的归一化图像处理后会产生一组热点图。每个热点反映了该像素作为特征点的概率大小。残差模块作为沙漏网络中的基础模块是本文的主要优化对象，其中，C代表连接，“＋”代表逐个元素的求和。本文提出的层次化并行模块（P-Residual Block，PRB）增加了跳转连接，提高了网络中的梯度流。同时该模块中并不存在1×1卷积，因为在二值网络中这种卷积核会造成比较严重的性能下降。

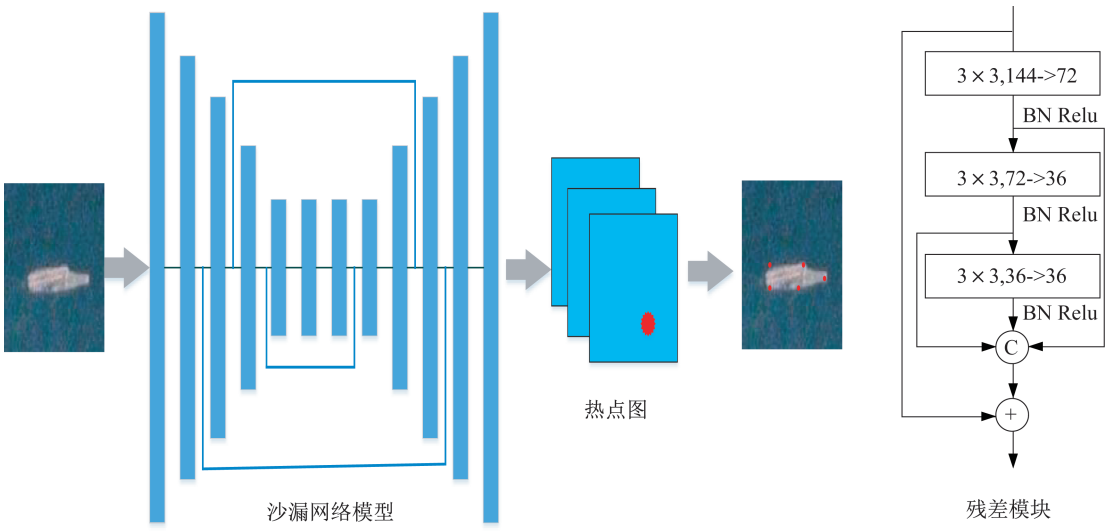

图 1　通用的沙漏网络模型架构及残差模块

Fig.1　Common hourglass network model and residual block

在PRB模块中,存在很多模块间与模块内的跳转连接。由于模块中每一层的激活值不能全部存在片上,从而造成大量的片外存储访问操作,并进一步限制网络推理速度并造成比较高的能耗开销。因此,PRB模块是限制整个B-HN网络性能的一个主要瓶颈。

此外,批量归一化_线性修正单元(Batch Normalization Rectified Linear Unit, BN_ReLU)被重新调整到卷积层之后,跳转连接之前来进行存储访问。为了进一步利用ReLU函数之后造成的稀疏性,采用多级预测校准(Multi-level Prediction-Correction, MPC)模型来降低大量零值激活值与非最大值之间的冗余操作。

B-HN的网络拓扑结构是对称的,任务精度与速度可以通过调整堆叠的沙漏模块数量来实现灵活调整。同时,通过调整B-HN的权重和最后一层的输出通道数来实现不同的检测功能。因此,本文提出的方案可以根据不同的任务需求(处理速度、功耗、识别精度等)对网络架构和参数进行灵活配置。

1.2　基于LUT方法的计算优化

考虑有限的二值权重可能会造成的大量重复的加法操作,本文提出了基于输入特征图重用的LUT方法来替换这些重复的操作。在B-HN网络中,需要强调的是第一层和最后一层的权重是多比特位宽。在B-HN网络的所有计算当中,多比特位宽权重卷积操作大概占比达到43%。因此,针对多比特位宽权重卷积操作进行优化将同样重要。

在文献[13]中,LUT可以被配置成2种模式:二值权重模式和多比特权重模式。一个深度为4的LUT可以执行针对多比特权重卷积的两路位串行加法操作和二值权重卷积的三路位串行加法操作。为了同时支持这2种模式,本文提出了一种统一模式的LUT方法。

1.2.1　二值权重卷积优化

对于二值权重,每个权重参数的值或者为−1或者为+1。如果将几个二值权重划分为一组,则这些权重的组合也是有限的。如图2所示,以M个输入特征图、N个输出特征图,以及滤波器尺寸为3×3为例,则一共需要$N\times M$个3×3的滤波器核。如果每3个权重划分为一组,对于相应的3个输入激活值a、b、c而言,一共有8种输出组合。在通常情况下,卷积神经网络中输出通道的数量要比这些组合的数量大很多,可以通过LUT的方法来避免重复操作。同时可以发现,LUT中出现的8种组合的值是对称的,即前4种组合可以通过相应的后4位组合直接取反得到。因此,只需要对4种组合进行编码即可以覆盖所有可能出现的组合,最后2 bit用来对相应的LUT值进行选择,第一比特则用来决定符号。

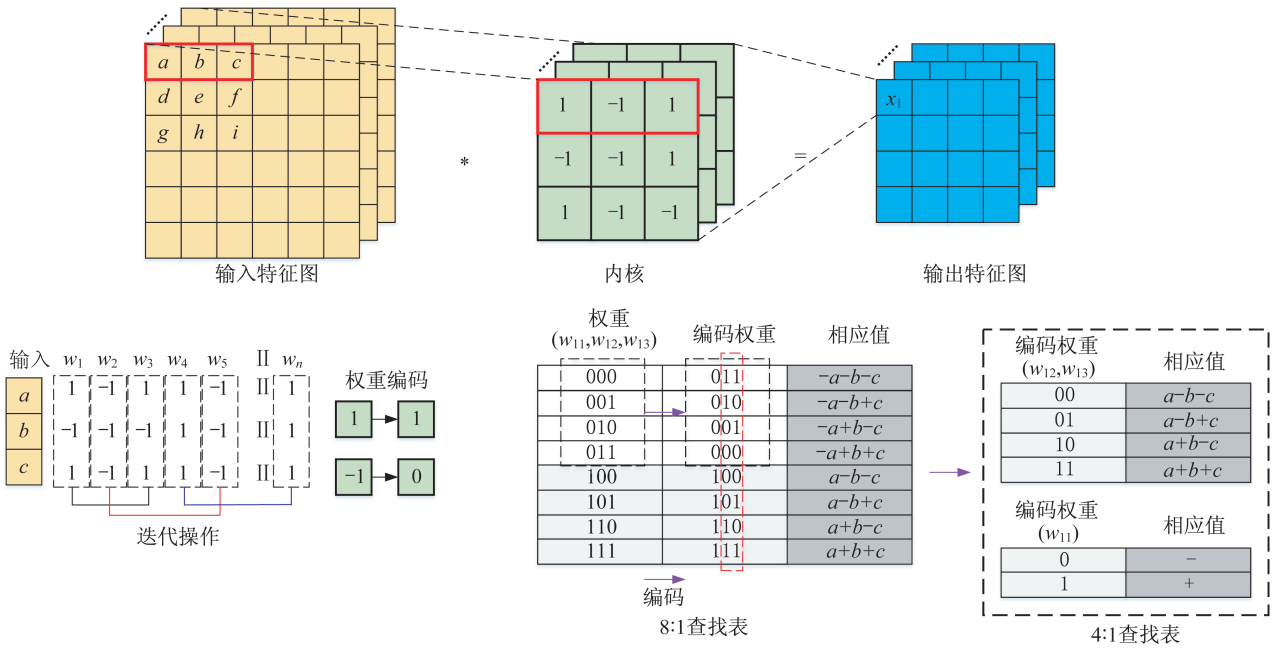

图 2　基于 LUT 的二值权重卷积计算优化

Fig.2　LUT-based binary weight convolution optimization

1.2.2　多比特卷积转换方法

考虑到二值权重计算的高效性，多比特权重也被转换成二值权重的形式，从而使得其乘法卷积转换成二值权重的计算方式。首先，B-HN 中第一层和最后一层的权重首先被量化到 8 bit。8 bit 权重值又被进一步拆分成 8 bit 的基值和 8 bit 的校准值。所有卷积核中基值的每个 bit 都是－1，而校准值的每一位则可能是＋1，也可能是－1。在对应同样输出位置的不同通道间计算时，基于基值的卷积计算结果都是一样的，因此，仅需要计算一次。然后通过校准值的计算来复用二值计算模式得到最终的计算结果。

同样以 M 个输入特征图、N 个输出特征图、卷积核尺寸 $k\times k$ 的 I bit 权重为例，为了计算同一位置所有输出通道的值，需要考虑 3 个方面的开销：所有权重基础组合的卷积计算、部分和的加法以及 LUT 值的选择。假设每 s 个权重为一组，并通过 P_1、P_2 和 P_3 来表示上述 3 方面的功耗开销。我们的目标是使得总的功耗开销最小化。需要强调的是，s 是正整数。通过解这个优化问题，可以得到最小开销下的 s 值。s 主要与输出通道数有关，基于 s 的加法操作降低比例见表 1，降低比率随着输出通道数的增加而快速增加。

表 1　不同输出通道数量下的计算优化对比

Tab.1　Computation optimization for different numbers of output channels

输出通道数	3～8	9～90	＞90
每组权重/bit	2	3	4
降低比例/%	16.67～37.50	44.44～64.44	71.15～75.00

另一个需要解决的问题是如何对权重进行分组。通过实验发现，经过 ReLU 函数处理后，统一通道的输出特征图中存在着大量连续分布的零值。这就意味着可以通过时钟门控来实现大量零值计算的跳转，因此相同通道内的权重被选择作为一组。与相关 CNN 加速器的 LUT 方法对比，本文提出的 LUT 计算方法可以实现对多种位宽权重的卷积实现统一支持，同时降低硬件的实现开销，针对多比特权重卷积实现 1.33～1.50 倍的卷积速度提升。

1.3　多级预测校准模型

在主流 CNN 模型中，通用的基础网络架构为“卷积-归一化-ReLU 激活-池化”等。其中，作为应用最广泛的激活函数，ReLU 会造成很高的激活值稀疏度。同时经过最大值池化层后，仅需要决定池化结果的相对尺寸，这也就意味着最终结果为负值

或非最大值的全精度卷积计算是完全没有必要的。与当前提出的基于零值的操作消除方法[13-14]相比，本文提出了一种更加高效的多级预测较准方法。

为了便于硬件实现，计算过程中的所有权重及中间结果均需要进行定点化。本文提出的多级预测校准主要针对两种计算模式：1) 激活层。预测阶段主要对最终激活值的符号进行判断，后面对结果进行校准。2) 最大值池化层。前面的计算判定激活值的相对尺寸，后面的计算完成更大值的更新。而在本文的B-HN网络中主要包括二值权重和多位宽权重两种网络类型。

在二值权重网络层，由于权重是1 bit数据，卷积操作被转换为加法操作。针对输入的8 bit数据，将其划分为高4 bit和低4 bit两个部分。对于激活层而言，在预测阶段，将高4 bit参与到计算过程并判断ReLU最终结果的符号，如果可以判定结果为负值，则将输出结果直接设置为0，后续结果无需继续执行；否则，需要执行低4 bit计算来对输出结果进行校准。在B-HN网络中，经过ReLU处理后的稀疏度在30%到90%之间，因此，采用本方法大概可以节省15%～45%的操作。对于最大值池化层，在预测阶段，首先计算高4 bit的结果来判断最大池化输出的相对尺寸。需要采用一个阈值T_m来对相对尺寸进行评估。如果最大值与其他非极大值的距离大于阈值，则后续非计算值的计算不需要继续执行。否则，需要执行后续计算过程。在校准阶段，低4 bit的计算结果将用来继续执行比较，本方法可以降低大约37.5%的计算处理。

在多比特权重网络中，输入特征值和权重均被量化为8 bit。对于输入而言，同样被分为高4 bit和低4 bit两部分，而权重则按每2 bit分组，从高位到低位划分为4部分。在此划分模式下，每次输入与权重的乘法操作转换为8级从高位到低位的计算过程，具体的计算及校准过程与二值权重计算过程相同。

2 硬件架构设计

如图3所示，B-HN加速器的硬件架构中主要包括3个模块：1) 预处理模块；2) 多级计算控制模块；3) 存储系统。

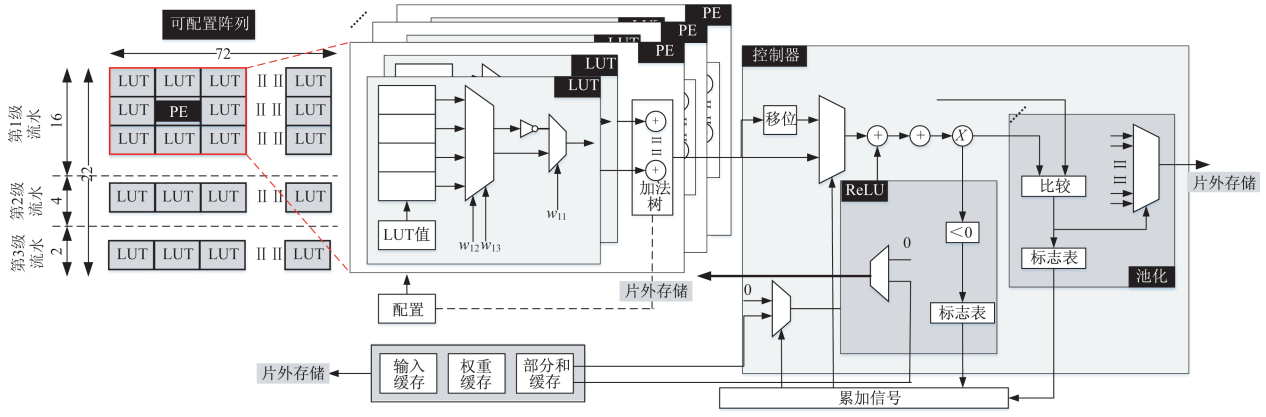

图3 B-HN加速器的系统架构

Fig.3 System architecture of B-HN accelerator

2.1 预处理模块

在B-HN加速器的处理过程中，输入的激活值及首末层权重的位宽均为8 bit，其他层的权重为1 bit。处理单元(Processing Element，PE)主要由查找表和加法器树组成。每个PE包括9个查找表和1个由9个加法器组成的加法器树，其中，每个LUT中包括4个7 bit的实体。在该架构中，总共包括176个PE。

加法器在配置下首先被配置来计算LUT中的4个基础值，编码后权重用来决定Psum的值。每组权重包括3 bit，由后2个bit来选择LUT中的值，第一个bit来确定符号。随后Psum值在加法器树中进行累加。最终，通过权重转换得到的校准值与Psum相加得到最后的结果。在整个处理过程中，二值权重卷积操作可以在一个周期内完成，而多比特权重操作则需要多个周期穿行计算完成。

为了优化片上数据存储及访问，采用了同时复用输入和权重的计算模式。对于多比特位宽层，输入尺寸为256×256×3（位宽为8 bit），在重用输入与权重的基础上利用所有的PE来并行计算3×256×3大小的模块，可以在每8个周期内完成64个点的计算。

而对于其他层而言，PE用来并行的执行2×16×144大小的模块，每16个周期产生32个点。对于残差模块，分别采用128、32和16个PE来完成残差模块中3层的计算，每6个时钟周期产生4个点。在整个计算过程中，所有的输入会一致被重用直到相对应通道的计算全部完成。

2.2　多级计算控制

该模块的主要功能是完成所有的参数配置和模块工作调度，包括阈值设定、多级计算预测、Psum累加、多比特权重卷积的分解与组合。该模块主要包括4个部分：归一化模块、ReLU模块、最大池化模块和确认模块。归一化模块包括一个乘法器和加法器。

ReLU模块则采用符号位来作为选择信号来判定选择原输出值或者是0。ReLU之后结果的符号位则写进标志位表来标识是否需要执行下一级计算以及控制PE阵列的数据更新，最大池化模块则由一个阈值比较逻辑和选择最大值的多选器组成。

2.3　存储系统

采用独立的输入缓存和权重缓存来为PE阵列提供数据。由于有限的数据带宽及计算资源，无法将卷积操作完全并行展开执行。因此，多级计算需要串行执行并且在片上实现每级计算的中间结果存储，采用Psum缓存来存储这些数据。对于残差模块中的跳转连接而言，部分数据需要存在很长时间才会被用于计算，因此，需要一个临时数据缓存来存储这些数据。由于本文提出的BCP方法，使得这些数据不需要存到片外，而仅需要存3行特征图数据。因此对于片外存储的带宽和访问极大降低。

2.4　基于模块计算的流水架构

作为B-HG网络中的瓶颈模块，PRB模块主要由内部及外部跳转连接的3层卷积层组成。考虑到每个中间层的输出结果均非常多，如果采用传统逐层处理方式的话，会造成与片外存储之间的大量数据交互。因此，提出面向PRB模块的三级流水计算架构。在这种架构下，仅仅需要在片上存储3.5行特征图数据，而且中间结果不需要写回到片外存储，同时跳转连接也不需要从片外存储多次读取数据。PRB模块中3层网络的计算时间比为8∶2∶1，根据这个比例来对每一层分配相应的计算资源，来保证流水处理的负载均衡性。

假设M个$W\times H$大小的输入特征图，N个输出特征图，卷积核尺寸为$k\times k$，将计算模块的尺寸设置为$w\times h$，输入、权重及中间结果的位宽分别为B_a、B_w和B_p。每次数据读取和写入的功耗分别为P_r和P_w，则由输入、位宽及中间结果数据访问造成的开销C_1、C_2和C_3分别为

$$C_1 = W \times H \times M \times B_a \times P_r \tag{1}$$

$$C_2 = \frac{W}{w-2} \times \frac{H}{h-2} \times M \times N \times k^2 \times B_w \times P_r \tag{2}$$

$$C_3 = \frac{M}{c} \times N \times W \times H \times B_p \times \left(P_r + P_w\right) \tag{3}$$

本文的设计目标则是实现三者之和的最小化，在给定的硬件资源约束下对这一优化问题进行求解，得到对于PRB模块每一层最优的计算模块尺寸为6×4×48、4×4×24和4×4×12。

对于非PRB的网络层而言，采用逐层处理的方式。为了保证最小的存储开销，对于每一层选取最优的模块尺寸。对于B-HN网络的第一层，其卷积核尺寸为3×3，步长为2，输出通道数为144，每次读取5行图像数据，计算模块被配置成3×128×3。对于B-HN网络中的其他网络层，卷积核尺寸为1×1，步长为1，计算模块尺寸则被配置成2×16×144。

3　性能评估与硬件实现

3.1　性能评估

为了进一步衡量本文提出方法的通用性，以AlexNet[15]作为测试基准来对本文提出的方法效果进行测试，网络的权重根据本文方法确定为每4 bit为一组。二值权重AlexNet网络中5层卷积层的性能统计数据见表2。表中可见，二值权重AlexNet的计算量可以降低82.75%。

表 2 AlexNet网络的性能优化分析

Tab.2 Performance optimization analysis for AlexNet

卷积层	输出通道数	初始操作数量/$\times10^9$	激活后零值比例/%	优化后操作数量/$\times10^9$	降低比例/%
Conv1	96	0.42	38.70	0.097	76.90
Conv2	256	0.90	72.50	0.151	83.19
Conv3	384	0.60	79.30	0.094	84.36
Conv4	384	0.45	77.60	0.071	84.14
Conv5	256	0.30	80.77	0.047	84.28
Total	—	2.67	—	0.460	82.75

3.2 硬件实现

在TSMC 28 nm的CMOS工艺下对提出的B-HN加速器进行了后端实现和性能仿真,具体的芯片版图及性能指标如图4和表3所示。

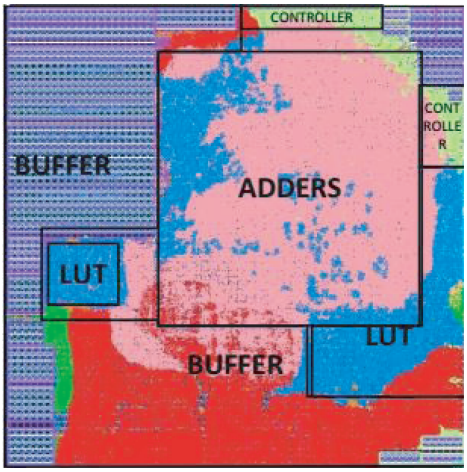

图 4 B-HN加速器后端版图

Fig.4 Post-layout of the B-HN accelerator

表 3 B-HN加速器后端仿真性能数据

Tab.3 Performance data of simulation for the B-HN accelerator post-layout

性能指标	本文结果
工艺节点/nm	TSMC 28 nm 1P10M CMOS
频率/MHz	500.0
片上存储/kB	92.0
面积/mm^2	0.7
吞吐率/(帧·s^{-1})	450.0
功耗/mW	117.0

B-HN加速器与相关二值权重硬件加速器的性能对比情况见表4。表中可见,B-HN加速器在计算延迟及功耗效率方面较相关工作取得了比较明显的提升。

表 4 与相关工作性能对比

Tab.4 Performance comparison with related works

性能指标	Yoda NN[16]	ISSCC 2018[8]	本文
工艺节点/nm	65.0	65.00	28.00
面积/mm^2	1.9	16.00	0.70
电压/V	1.2	1.10	0.90
频率/MHz	480.0	90.00	500.00
延迟/ms	71.0	30.00	2.20
功耗/mW	40.0	31.20	117.00
峰值性能/GOPS	1 510	19.70	1 188
功耗效率/(TOPS·W^{-1})	2.2	0.92	10.15

4 结束语

本文首先采用二值化权重方法对当前通用的沙漏卷积神经网络模型进行处理,并在进一步分析其计算、存储瓶颈的基础上提出了基于多级预测校准模型及LUT方法的高效卷积计算、针对残差模块的基于模块计算的流水架构,最终在28 nm工艺条件下对提出的B-HN模型高效硬件设计及性能评估。后续工作将重点基于航天应用中的专用数据集本文架构进行进一步优化。

参考文献

[1] 刘鑫博.基于FPGA的卫星图像在轨处理技术研究[D].哈尔滨:哈尔滨工程大学:2016.

[2] 袁秋壮,魏松杰,罗娜.基于深度学习神经网络的SAR星上目标识别系统研究[J].上海航天,2017,34(5):46-53.

[3] 周舟,王海鹏,徐丰,等.基于通道剪枝的SAR图像舰船检测优化算法[J].上海航天,2020,37(4):48-54.

[4] 石国强,赵霞,陈星洲,等.基于卷积神经网络的局部图像特征描述符算法[J].上海航天,2020,37(1):91-96.

[5] 周敏,史振威,丁火平.遥感图像飞机目标分类的卷积神经网络方法[J].中国图象图形学报,2017,22(5):702-708.

[6] 李亚飞,董红斌.基于卷积神经网络的遥感图像分类研究[J].智能系统学报,2018,13(4):550-556.
[7] 郭倩,王海鹏,徐丰.星载合成孔径雷达图像的飞机目标检测[J].上海航天,2018,35(6):57-64.
[8] LEE J, KIM C, KANG S, et al. UNPU: a 50.6 TOPS/W unified deep neural network accelerator with 1b-to-16b fully-variable weight bit-precision [C]// IEEE International Solid-State Circuits Conference. 2018: 218-220.
[9] KIM H, SIM J, CHOI Y, et al. A kernel decomposition architecture for binary-weight convolutional neural networks [C]// ACM Annual Design Automation Conference. 2017: 1-6.
[10] ZHENG S, LIU Y, YIN S, et al. An efficient kernel transformation architecture for binary-and ternary-weight neural network inference [C]// ACM Annual Design Automation Conference. 2018: 1-6.
[11] LU W, YAN G, LI J, et al. Flex flow: a flexible dataflow accelerator architecture for convolutional neural networks [C]// IEEE International Symposium on High Performance Computer Architecture. 2017: 553-564.
[12] ALWANI M, CHEN H, FERDMAN M, et al. Fused-layer CNN accelerators [C]// IEEE International Symposium on Microarchitecture (MICRO). 2016: 1-12.
[13] AKHLAGHI V, YAZDANBAKHSH A, SAMADI K, et al. Snapea: predictive early activation for reducing computation in deep convolutional neural networks[C]// IEEE Annual International Symposium on Computer Architecture (ISCA). 2018: 662-673.
[14] SONG M, ZHAO J, HU Y, et al. Prediction based execution on deep neural networks[C]// IEEE Annual International Symposium on Computer Architecture (ISCA). 2018: 752-763.
[15] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60(6): 84-90.
[16] ANDRI R, CAVIGELLI L, ROSSI D, et al. YodaNN: an architecture for ultralow power binary-weight CNN acceleration [J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2017, 37(1): 48-60.

(上接第 51 页)

[18] SCHUSTER M, PALIWAL K K. Bidirectional recurrent neural networks [J]. IEEE Transactions on Signal Processing, 1997, 45(11): 2673-2681.
[19] CHEN T, GUESTRIN C. Xgboost: a scalable tree boosting system [C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2016: 785-794.
[20] HADDAD N, BROWN R, FERGUSON R, et al. SOI: is it the solution to commercial product SEU sensitivity? [C]// ESA Special Publication. Paris, France: ESA, 2004: 231.