

引用格式:李永博,王琴,蒋剑飞. 稀疏卷积神经网络加速器设计[J]. 微电子学与计算机, 2020, 37(6): 30-34. [LI Y B, WANG Q, JIANG J F. Design of sparse convolutional neural network accelerator [J]. Microelectronics & Computer, 2020, 37(6): 30-34.]

稀疏卷积神经网络加速器设计

李永博,王 琴,蒋剑飞

(上海交通大学 电子信息与电气工程学院,上海 200240)

摘 要: 为降低卷积神经网络推断时的时延和能耗,使用动态网络剪枝技术得到稀疏网络并设计出高能效比的稀疏卷积神经网络加速器. 针对运算负载不均衡问题,提出适合稀疏运算的数据流;针对卷积运算高时延问题,采用 16×16 运算阵列提高运算并行度,设计索引单元避免无效运算,设计脉动输入层加强数据复用,采用乒乓缓存减少数据等待. 综合结果表明,在 TSMC 28 nm 工艺下,芯片工作频率可达 500 MHz,功耗为 249.7 mW,卷积运算峰值算力达到 256 GOPS,能效比为 1.03 TOPS/W.

关键词: 稀疏卷积神经网络;阵列运算;加速器;高能效比

中图分类号: TN492

文献标识码: A

文章编号: 1000-7180(2020)06-0030-05

Design of sparse convolutional neural network accelerator

LI Yong-bo, WANG Qin, JIANG Jian-fei

(School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University,
Shanghai 200240, China)

Abstract: In order to reduce the latency and energy consumption of convolutional neural networks, dynamic network surgery is used to get sparse networks and a high energy efficiency sparse convolutional neural network accelerator is designed. Aiming at the problem of unbalanced computing load, a dataflow suitable for sparse computing is proposed. To reduce the latency of convolution operation, a 16×16 process engine array is used to improve computation parallelism, index units are designed to avoid invalid operation, the systolic input structure is designed to enhance data reuse, and ping-pong buffers are introduced to reduce data waiting. The synthesis results show that the frequency can reach 500 MHz, the power consumption is 139mW, the peak performance is 221 GOPS, and the energy efficiency is 1.59T OPS/W with TSMC 28nm process.

Key words: sparse convolutional neural network; array computation; accelerator; high energy efficiency

1 引言

卷积神经网络(Convolutional Neural Network, CNN)的发展表明,通过增加网络深度可以有效提高识别精度^[1],但更多的权重参数导致 CNN 推断时延长、功耗高、存储需求大,不利于边缘计算. 如何降低网络复杂度并设计出符合 CNN 运算模式的高效系统成为突破人工智能部署瓶颈的关键.

算法层面,文献[2]通过剪枝、量化和哈夫曼编码等技术,稀疏化网络并降低位宽,将模型大小缩减 35-49 倍. 但硬件层面,如何充分发挥稀疏 CNN 的优势成为一大难点. Eyeriss^[3]通过探索数据复用和功耗的关系,提出行不变数据流,实现低功耗加速. 但 Eyeriss 未能充分利用 CNN 稀疏性,乘数为 0 的运算无法被省略,运算时延没有降低. Cambricon-X^[4]通过探索 CNN 的稀疏性,在传统硬件加速器基

础上增添寻址模块,避免无效运算,提升了运算速度和运算效能,但其寻址模块较复杂,

引入了 31% 的额外功耗. 本文使用动态网络剪枝技术得到稀疏 CNN;设计精简的寻址模块并探索适合稀疏运算的数据流;提出高效比加速器架构,使用 TSMC28nm 工艺综合并与同类研究相比较.

2 稀疏卷积神经网络及其优化设计

2.1 基本概念

CNN 通常由卷积层、池化层、激活函数层和全连接层构成. 其中卷积层是提取区域特征的主要手段,其经典结构如图 1 所示. C_{in} 个输入通道与 C_{out} 个深度为 C_{in} 的三维卷积核完成卷积累加运算,得到 C_{out} 个输出通道,添加偏置后进入激活层和池化层.

表 1 卷积层参数及其含义

卷积层参数	含义
C_{in}	输入通道个数
C_{out}	输出通道个数
N_{in}	输入特征图尺寸
N_{out}	输出特征图尺寸
K	卷积核尺寸

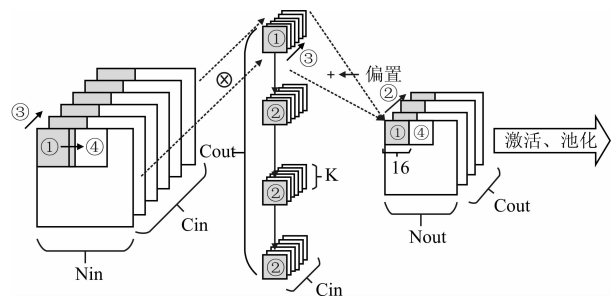


图 1 卷积层结构

卷积层是 CNN 运算量和参数量最大的部分,例如 AlexNet 中卷积层运算量占总运算量的 92%,卷积运算能力成为制约网络速度的最主要因素. 因此大多数论文在加速 CNN 运算时,主要讨论对卷积层的优化,本文也着重于实现高效比的稀疏卷积层运算.

2.2 动态网络剪枝及验证

CNN 模型参数量巨大,例如 VGG-16 拥有 16 层、1.38 亿个参数,这些参数的重要性不同,可将不重要的参数删除,通过重训练恢复精度. 剪枝后的卷积核大量权重为 0,表现为稀疏矩阵. 通过特殊设计

硬件,存储空间和带宽需求减小,运算周期和总能耗降低.

网络参数的重要性随着训练过程的迭代而变化,文献[2]直接删除参数会导致精度下降较多且训练缓慢,因此本文采用动态网络剪枝算法(Dynamic Network Surgery,DNS)实现稀疏化. 首先输入一个预训练的网络模型,然后根据设定的阈值,裁剪重要程度低的连接;通过反向传播更新网络参数,再次评估更新完后该模型各权重的重要性,将其中重要程度低的连接删除,将重要度高却被上一步删除的连接复原,不断迭代直至精度恢复.

本文使用 Tensorflow 平台基于 ImageNet 数据集,完成 DNS 技术的验证. MobileNet、VGG-16 和 AlexNet 等代表性网络的 Top-5 准确率变化如表 2 所示.

表 2 DNS 后准确率变化 %

神经网络	稀疏度	原始错误率	DNS 错误率	准确率变化
MobileNet	23.38	10.10	10.88	-0.78
VGG-16	74	11.35	12.54	-1.19
AlexNet	72	22.64	22.54	+0.10

三种网络的 Top-5 准确率变化在 1% 左右. VGG-16 与 AlexNet 网络权重较多,因此网络稀疏度可达到 70% 以上,即 70% 以上的权重为 0. MobileNet 是轻量化网络,冗余参数较少,网络稀疏度只能达到 23%. 因此,非轻量化 CNN 中,DNS 技术能够在精确度基本不变的前提下,得到高稀疏度网络.

2.3 批归一化层的融合

MobileNet 等网络引入批归一化(Batch Normalization,BN)层以解决训练时梯度消失和梯度爆炸. 推断时 BN 层表现为在卷积层后增加一层常数运算,如公式(1)(2)所示. 其中 σ^2 为方差, μ 为均值, β 为偏移量, γ 为缩放因子,四者在推断时为常数, ϵ 为较小数防止分母为 0. X 为输入图像, W 为权重, B 为偏置, Y_{conv} 为卷积层输出, Y_{bn} 为 BN 层输出.

$$Y_{conv} = W \times X + B \tag{1}$$

$$Y_{bn} = \gamma \times \left(\frac{Y_{conv} - \mu}{\sqrt{\sigma^2 + \epsilon}} \right) + \beta \tag{2}$$

将 Y_{conv} 代入 Y_{bn} 中并化简,可以得到 BN 层与卷积层融合后的计算公式,其运算模式与融合前的卷积层相同,权重与偏置发生了改变,如公式(3) - (5) 所示.

$$Y_{bn} = \hat{W} \times X + \hat{B} \quad (3)$$

$$\hat{W} = \frac{\gamma}{\sqrt{\sigma^2 + \epsilon}} \times W \quad (4)$$

$$\hat{B} = \frac{\gamma}{\sqrt{\sigma^2 + \epsilon}} \times (B - \mu) + \beta \quad (5)$$

融合 BN 层前后的网络,推断时在数学上等价,且权重为 0 的连接乘以变换因子后仍为 0,因此融合 BN 层不影响网络的准确度和稀疏度,同时能够降低运算量。

3 CNN 加速器 ASIC 实现

3.1 并行数据流设计

图 1 所示的卷积层结构,在算法上可抽象为串行执行的 6 层循环。本文将循环分为四类:输出通道遍历(L_Cout)、输入通道遍历(L_Cin)、卷积窗滑动(L_Nout)和卷积运算(L_K)。根据循环展开方式不同,并行方式可分为输入通道并行(L_Cin 展开)、输出通道并行(L_Cout 展开)、多卷积窗并行(L_Nout 展开)和卷积运算并行(L_K 展开)。

稀疏 CNN 中,卷积运算并行会导致 PE 空置。例如 $1 \times 3 \times 3$ 卷积核分配至 9 个 PE 运算,当稀疏度大于 0% 时,有 PE 无运算数输入,导致 PE 利用率降低。

稀疏 CNN 不同卷积核稀疏度不同,输入、输出通道并行会导致 PE 阵列运算负载不均衡,相互等待。输入通道并行,不同输入通道的卷积结果到达累加模块的时刻不同,导致先算完的 PE 发生等待;输出通道并行,同一输入通道的不同卷积核运算完成的时刻不同,在不同时刻请求切换输入通道,然而输入通道需等到最慢的卷积核运算完成后才能切换,也会导致先算完的 PE 发生等待,降低 PE 利用率。

对卷积核预排序,即将稀疏度相似的卷积运算并行,可以改善 PE 等待的情况。但所有卷积运算不能任意排序,需要保证 M 个输出通道各通道对应的卷积运算属于相同的 N 个输入通道,以便累加。对 VGG-16 预排序的效果如图 2 所示。预排序后的理论运算周期数(不考虑 DDR 访存时间、图像传输时间等)最高减少约 45% (90% 稀疏度下),比预排序前提升约 20%,但无法达到理论最优值。

多卷积窗并行的数据流可以解决 PE 相互等待的问题。这种数据流下,加速器的所有 PE 计算同一个输入和输出通道的不同位置的卷积窗,此时所有

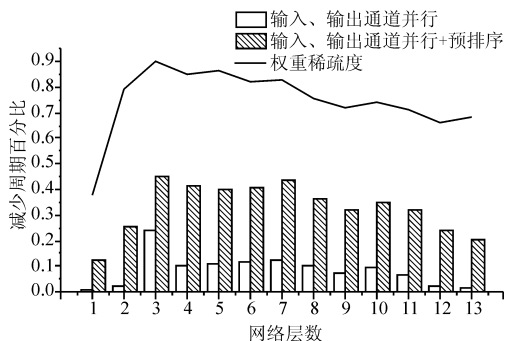


图 2 输入、输出并行时理论加速比与稀疏度比较

PE 计算的卷积核相同,呈现同步的状态,理论加速比可以达到该层的稀疏度。

为保证稀疏 CNN 运算时 PE 的利用率,本文选择多卷积窗并行的数据流,其他循环的并行度设为 1。具体数据流如图 1 所示。①部分输入图像输入并存储在片上, 16×16 PE 阵列并行完成该图像某一个卷积核同一通道的 16×16 个卷积窗的运算,得到 16×16 的中间值特征图;②遍历输出通道并重复①中的运算;③输出通道遍历完成或片上存储满后,切换输入通道,重复①②的运算,完成中间值图像的累加;④输入通道遍历后,得到部分特征图的最终结果并输出,输入框滑动,不断重复①②③④中的运算直至卷积层运算完毕。

3.2 硬件结构设计

3.2.1 ASIC 整体架构

本文设计的硬件加速器包含脉动输入模块、 16×16 PE 阵列、权重缓存、中间值缓存、输出仲裁和控制模块。其中脉动输入模块以脉动的形式向 PE 阵列输入特征图, PE 阵列主要负责权重索引、卷积运算和中间值累加,权重缓存负责存储和广播权重,中间值缓存负责存储中间值,输出仲裁负责输出最终结果,控制模块负责配置 PE 阵列。其中,运算阵列可分为 16×16 、 14×14 、 $4 \times 8 \times 8$ 等多种运行模式,以更加高效地完成卷积运算。当输出图像较大时,全部 16×16 个 PE 参与运算。当输出图像用 14×14 与 16×16 个 PE 运算迭代次数相同时,以 14×14 阵列模式运算。在深层卷积层中,输出图像的尺寸可能小于等于 8×8 ,若只启用有效 PE, PE 利用率低于 25%。本文将运算阵列拆分为 4 个 8×8 的子阵列,每个子阵列有自己的权重缓存,使其能够并行运算 4 个不同输出通道,提高 PE 理论利用率。

卷积运算时,控制模块首先将阵列配置为适合本层卷积运算的模式。之后数据从片外输入,其中图

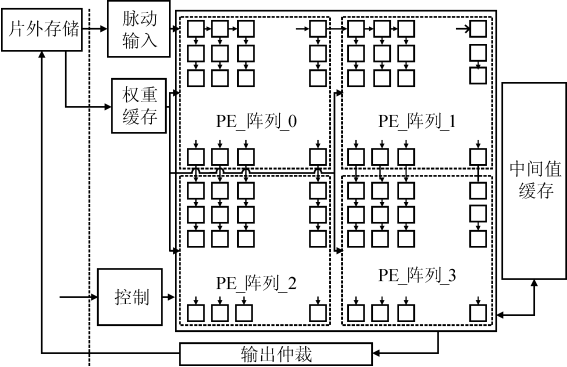


图 3 ASIC 整体架构

像数据以脉动方式输入至 PE 阵列,权重输入至权重缓存.权重缓存广播权重标志着运算开始,经过索引和乘累加后得到中间值,并存储在中间值缓存中.遍历输入通道时,中间值从缓存中进入 PE 阵列参与累加,得到最终结果.最终结果进入输出仲裁等待输出,同时切换图像及权重数据,重复上述操作直至一层卷积层运算完毕.

3.2.2 PE 设计

PE 阵列中每个 PE 的结构相同,如图 4 所示.输入图像经过索引模块得到图像有效值,该值与权重有效值进入乘累加单元完成卷积运算,之后与中间值缓存中存储的值进行不同输入通道的累加,累加结果存回中间值缓存的相同地址中,不断迭代直至得到卷积层输出.

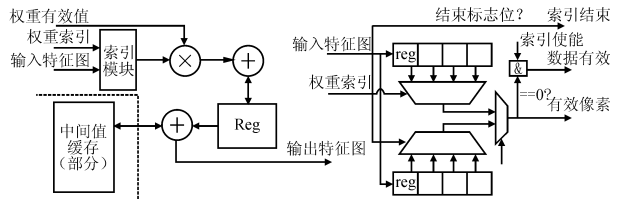


图 4 PE 结构(左)及索引模块结构(右)

不同于 Cambricon-X^[4] 存储权重步长,再在片上计算实际索引地址的方法,本文的权重直接以{实际索引地址,有效值}的形式存储,如图 5 所示,稀疏卷积核直接存储为{0,W0}{3,W3}.运算时,索引用于取出对应地址的像素,即索引 0 取出像素 I0,索引 3 取出像素 I4,有效像素进入运算单元与权重有效值进行乘累加.这种索引方式使 0 值权重不进入片上,最大程度发挥稀疏 CNN 的优势,且索引模块设计简单,功耗较低.图 4 所示实际电路中,Ping-Pong Buffer 用于隐藏输入图像传输时延,权重索引由权重缓存以广播的方式传入,通过 MUX 选出像素有效值.另外,若选出的像素值为 0,则将使能信号置 0

避免此次运算以降低功耗.

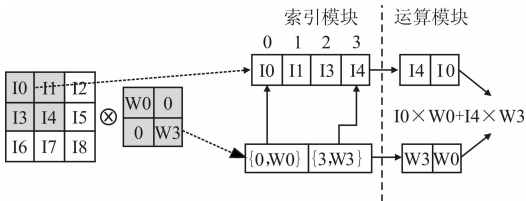


图 5 权重存储方式及索引模块原理

3.2.3 数据传输层设计

如何将数据快速且不重复地发送至运算单元是阵列架构需要探讨的关键问题之一.若 PE 与数据缓存全相联,传输速度快但布线资源紧张、数据重复发送较多;若总线形式相连,数据复用度高、开销小但传输带宽低.本节提出一种脉动传输方式,数据复用高且连线简单,同时将传输层与运算层解耦以隐藏传输时间.

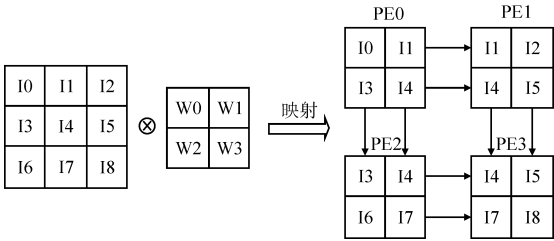


图 6 PE 阵列中输入数据的分布

Nin 为 3,K 为 2 的卷积运算映射到 2×2 PE 阵列中的数据分布如图 6 所示,在同一行 PE 中,一个数据可能被当前及其右方 PE 所需求.在整个阵列中,一行 PE 中的数据可能被当前及其下方 PE 所需求.基于上述数据分布,本文提出了脉动传输方式,如图 7 所示.数据从片外传入后,首先进入第一行 PE 中,向右移动直到每个 PE 得到其需要的数据.非第一行 PE 需要的数据,也需要先输入至第一行 PE,到达对应的列后一级一级向下转发,每行 PE 留下需要的数据.这样相关数据输入一遍,便可以灌满整个 PE 阵列,且对布线资源的需求较低.

上述脉动输入的缺点是更换输入图像时,PE 阵列需要较长时间才能填满.观察数据流发现,两次输入通道切换有较长的时间间隔用于遍历输出通道,且运算时间往往长于数据传输时间.为了避免运算单元发生等待,本文将传输与运算解耦合,每个 PE 都有专用于数据传输的寄存器,这样在运算时也可以传输数据.有效数据由传输寄存器存入该 PE 索引模块的乒乓缓存中,在切换输入通道后可立即参与运算,隐藏了数据传输时延.

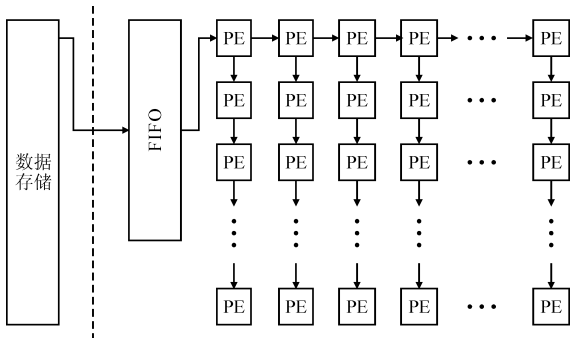


图 7 数据传输层示意

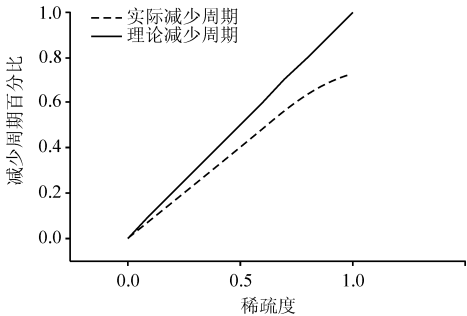


图 8 不同稀疏度下加速比

4 实验结果与分析

4.1 实验环境

本文实现的硬件架构包含 256 个 PE, 72 KB SRAM. 通过增量网络量化将乘法替换为移位后, 权重降低至 4 位. 该架构的功能和性能仿真通过 Synopsys VCS 实现, 测试集为 VGG-16. 使用 Synopsys Design Compiler, 基于 TSMC 28nm RVT 工艺对硬件设计进行综合, 并由 PrimeTime PX 仿真得到芯片功耗.

4.2 实验结果

本设计综合后最高频率达到 500 MHz, Floorplan 后总面积为 $1.5 \times 2 \text{ mm}^2$. 峰值算力达到 256 GOPS, 为了对比公平, 以稠密 VGG-16 模型为输入, 平均功耗为 249.7 mw. VGG-16 网络卷积运算峰值算力达到 221 GOPS, 功耗为 139 mW.

在不同网络稀疏度下, 忽略 DDR 交换的时延, 本设计卷积运算的加速比如图 8 所示. 随着网络稀疏度的升高, 硬件加速比与理论加速比的差值逐渐增加, 主要原因有 (1) 为了对齐中间值特征图以便累加、简化 SRAM 访存控制, 全 0 的卷积运算也需要向片内传输一个 0 值, 导致实际稀疏度低于理论稀疏度, 且稀疏度越高, 全 0 卷积运算出现的概率越大; (2) 切换输出通道时, 运算需暂停一个周期用于判断标志位. 稀疏度越高, 运算周期越少, 停滞周期占比越高, 导致加速比降低.

表 3 与同类研究比较结果

	本文	文献[4]	文献[6]	文献[7]
工艺(nm)	28	65	28	28
功耗(mW)	249.7	954	7.5 - 300	155
面积(mm ²)	3	6.38	1.87	6.3
频率(MHz)	500	1000	200	500
峰值算力(GOPS)	256	544	102	128
峰值能效比(TOPS/W)	1.03	0.57	0.3-13.6	0.83

为评估硬件架构的性能, 本文选取了同样加速稀疏 CNN 的 Cambricon-X, 同样工艺实现的 ASIC

以及 CPU、GPU 等通用处理器作为对比.

表 4 与通用处理器比较结果

	本文	i7-5930K	GeForce Titan X	Tegra K1
工艺(nm)	28	22	28	28
面积(mm ²)	3	356	601	—
频率(MHz)	500	3 500	1075	852
功耗(W)	0.25	73	159	5.1
GOPS	256	255.7	4 500	383.3
GOPS/mm ²	85.33	0.72	7.49	—

对比发现, 本文提出的架构功耗较低、能效比较高, 单位面积算力优于通用处理器.

5 结束语

本文在对剪枝算法研究的基础上, 进行了稀疏 CNN 的 ASIC 加速器设计. 通过分析不同并行方式对 PE 阵列利用率及推断速度的影响, 得到一种运算负载均衡的数据流; 提出 16×16 PE 阵列为主的硬件架构, 提高了卷积运算并行度; 设计多种阵列运算模式, 提高了 PE 利用率; 根据数据分布特性提出脉动式且运算与传输解耦的输入方式, 隐藏了传输时延; 设计索引模块避免无效运算, 充分利用了 CNN 稀疏性. 经过仿真与综合, 本设计的频率达到 500 MHz, 卷积运算理论峰值算力达到 256 GOPS, 功耗为 249.7 mW, 能效比达到 1.03 TOPS/W.

参考文献:

[1] WANG Q, SHEN F, SHEN L, et al. Lung Nodule Detection in CT Images Using a Raw Patch-Based Convolutional Neural Network[J]. Journal of digital imaging, 2019, 32(6): 971-979.

[2] HAN S, MAO H, DALLY W J. Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding[C]// International Conference on Learning Representations. San Juan, Puerto Rico, 2016: 1-14.

10B 编码器,采用 Verilog 语言实现了编码器的功能,通过 NCVerilog 与 Design Compiler 软件完成了对电路进行仿真与综合,并已应用于 12.5 Gbps 传输速率的 ASIC 芯片中.因此,本设计对安全可靠、自主可控的国防军工产品生产有较大的实用价值与借鉴意义.

参考文献:

[1] 常红,柯导明,孟坚,等. 新型 8B/10B 编码方案的设计与实现[J]. 计算机工程与应用,2018(2):87-90.

[2] 李长庆,程军,李梁,等. 采用并行 8b/10b 编码的 JESD204B 接口发送端电路设计[J]. 微电子学与计算机,2017,34(8):70-75.

[3] 王方,周璐,张正璠. 8B/10B 编码器新型算法结构的设计与实现[J]. 微电子学与计算机,2016, 33(10): 151-154.

[4] 霍兴华,姚亚峰,贾茜茜,等. JESD204B 接口协议中的 8B10B 编码器设计[J]. 电子器件,2015,38(5):55-59.

[5] WIDMER A X,FRANASZEK P A. A DC-Balanced, Partitioned-Block,8B/10B Transmission Code[J]. IBM

Journal of Research and Development, 1983, 27(5): 440-451.

[6] YADAV S,PANDEY S,GUPTA A. Implementation of 8B/10B encoder-decoder for Gigabit Ethernet Frame [C]//2014 Eleventh International Conference on Wireless and Optical Communications Networks (WOCN). IEEE,2014:1-4.

作者简介:

王俊杰 男,(1996-),硕士研究生,研究方向为大规模集成电路设计.

万书芹(通讯作者) 女,(1977-),博士,高级工程师,研究方向为大规模集成电路设计.
E-mail: wsqwx@aliyun. cn

季惠才 男,(1972-),研究员,研究方向为大规模集成电路设计.

陶建中 男,(1962-),研究员,硕士生导师,研究方向为大规模集成电路设计.

杨 阳 男,(1993-),硕士,研究方向为大规模集成电路设计.

(上接第 34 页)

[3] CHEN Y H, KRISHNA T, EMER J S, et al. Eyerriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks[J]. IEEE Journal of Solid-State Circuits, 2017, 52(1): 127-138.

[4] ZHANG S, DU Z, ZHANG L, et al. Cambricon-X: An accelerator for sparse neural networks[C]// 2016 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO). Taipei, China: IEEE, 2016:1-12.

[5] ZHOU A, YAO A, GUO Y, et al. Incremental network quantization: Towards lossless cnns with low-precision weights [J]. arXiv preprint arXiv: 1702. 03044, 2017.

[6] MOONS B, UYTTERHOEVEN R, DEHAENE W, et al. 14.5 Envision: A 0.26-to-10TOPS/W subword-parallel dynamic-voltage-accuracy-frequency-scalable Convolutional Neural Network processor in 28nm FD-

SOI[C]// Solid-state Circuits Conference. San Francisco, USA: IEEE, 2017: 246-247.

[7] ALESSANDRO A, HESHAM M, ENRICO C, et al. NullHop: A Flexible Convolutional Neural Network Accelerator Based on Sparse Representations of Feature Maps[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(3):644-656.

作者简介:

李永博 男,(1995-),硕士研究生.研究方向为神经网络加速器设计、大规模集成电路设计.
E-mail: lyb_sd@sjtu. edu. cn

王 琴 女,(1975-),副教授.研究方向为大规模集成电路设计、先进集成电路设计方法.

蒋剑飞 男,(1979-),助理研究员.研究方向为高速数字集成电路设计.