

引用格式:李倍,闵丰,杨军,等.一种基于深度学习的目标跟踪加速器[J].微电子学与计算机,2021,38(8):53-58. [LI B, MIN F, YANG J, et al. A deep learning object tracking accelerator[J]. Microelectronics & Computer, 2021, 38(8): 53-58.]

一种基于深度学习的目标跟踪加速器

李 倍¹, 闵 丰², 杨 军¹, 梁 科¹, 李国峰¹

(1 南开大学 IC 设计与系统集成实验室, 天津市光电传感器与传感网络技术重点实验室 天津 300350

2 中国科学院 计算技术研究所 北京 100190)

摘 要: 针对当前神经网络加速器难以高效实现目标跟踪边框后处理的问题, 提出一种高效的目标跟踪专用加速器. 引入神经网络架构, 用于提取输入视图特征并生成边框置信度与偏移量集合. 随后针对目标跟踪的边框处理设计了专用于边框的回归、惩罚以及提取操作的加速模块. 通过同步神经网络加速器与专用加速模块间的数据, 以流水结构并行执行特征提取与边框操作, 实现基于深度学习目标跟踪的端到端处理. 该加速器在 40 nm 工艺下消耗面积 3.64 mm², 获得了 5.71 Tops/W 能效比. 实验结果表明: 与现有加速方案相比, 该目标跟踪加速器获得了 1.53 倍加速, 可实现实时的视频处理(31 fps). 其中仅针对跟踪过程的后处理任务, 专用加速模块相对 RISC 处理器可实现 3.2 倍的加速比.

关键词: 深度学习目标跟踪; 后处理; 专用模块; 硬件加速器

中图分类号: TP183

文献标识码: A

文章编号: 1000-7180(2021)08-0053-06

A deep learning object tracking accelerator

LI Bei¹, MIN Feng², YANG Jun¹, LIANG Ke¹, LI Guofeng¹

(1. Tianjin Key Laboratory of Optoelectronic Sensor and Sensing Network Technology, Integrated Circuit and System Integration Laboratory of Nankai University, Tianjin 300350, China

2. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: Since the current neural network accelerator couldn't efficiently accelerate the post-processing of object tracking, a dedicated object tracker is proposed. A neural network architecture is introduced to extract the features of the input feature map. At the meanwhile, it generates the bounding box confidence and position offset sets. A dedicated acceleration module is designed for the anchor regression, penalty calculation and extraction. By synchronizing the data between the neural network accelerator and the dedicated module, a new pipelined structure is proposed to execute the feature extraction and anchor regression in parallel. Therefore, the end-to-end processing of the object tracking is efficiently achieved. The accelerator consumes an area of 3.64 mm² under the SMIC 40nm process, and achieves 5.71 Tops/W energy efficiency. Experimental results show that, compared with the current acceleration solutions, the object tracking accelerator achieves 1.53 times acceleration, and it could realize real-time video processing(31fps). For the post-processing task of the tracking, the processing speeds of the proposed dedicated module is improved by 3.2 times than the RISC processor.

Key words: DNN object tracking; post-processing; dedicated module; hardware accelerator

1 引言

目标跟踪是计算机视觉领域的基础研究之一,广泛地应用于自动驾驶、视频监控和机器人视觉导航等各个领域,因而视觉目标跟踪技术的研究有着重大的实用价值. 针对有些场景对目标跟踪的实施新需求^[1], 尽管当前的图形处理器 (Graphics Processing Unit, GPU) 能达到其速度要求, 但能耗与成本问题使其难以广泛应用到终端设备. 如一种在 GPU 上加速目标跟踪过程的方法^[2], 解决了现有技术跟踪性能不高和实时性差的问题, 但由于功耗大, 只能部署在云端服务器, 终端数据需要实时传输到云端服务器处理, 庞大的数据量导致云服务器的计算压力过大, 算力下沉至终端处理器. 近年来随着深度学习的发展, 基于该技术的智能跟踪方法随之兴

起, 并获得更优的效果, 然而当前的卷积神经网络复杂度高, 算力大, 如跟踪网络常用的前端算法 VGG-16 处理一帧图像需要 30.76 GOP, 因此新兴的目标跟踪算法的计算量也随之增加^[3], 当前的神经网络加速器难以满足其实时算力需求. 同时基于现有的神经网络加速器难以高效加速跟踪算法的边框处理过程, 阻碍了其目标网络的加速效率, 因而只能将边框处理部分的运算转移至 RISC 处理器, 但这将导致跟踪任务的系统延迟增大, 无法实现实时处理. 为了解决以上问题, 本文提出一种面向深度学习目标跟踪算法的硬件加速器, 该加速器包含一种深度流水处理的神经网络加速器, 结合一个独立设计的后处理模块组合为端对端流水架构, 可有效提升计算效率, 实现目标跟踪的实时处理.

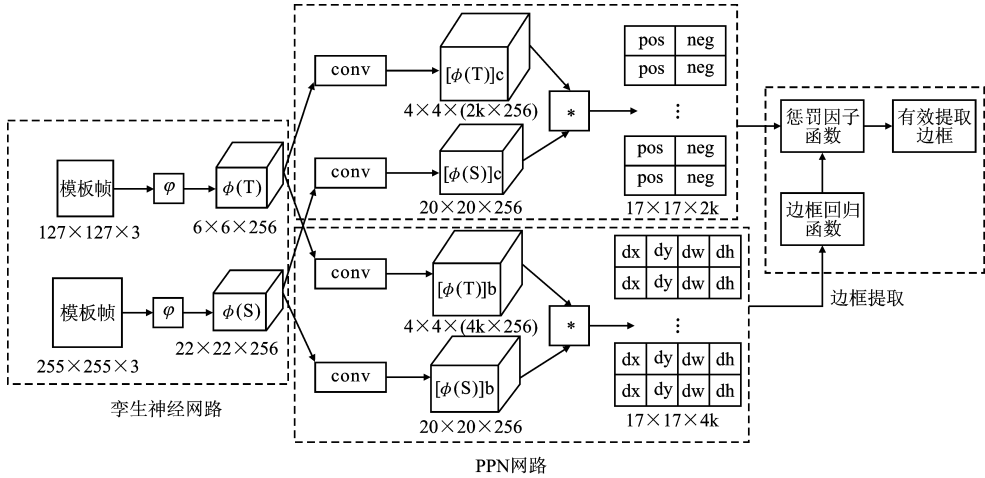


图 1 SiamRPN 算法框架图

2 目标跟踪器的整体方案

2.1 SiamRPN 跟踪算法

近年来, 深度学习目标跟踪算法以准确率高、稳定性强的特点迅速成为研究热点. 本文选用最具代表性的 SiamRPN 算法^[4] 进行加速设计, 其卷积部分为孪生神经网络, 后处理网络由区域提取网络 (Region Proposal Network, RPN) 和边框提取单元构成, 整体框图如图 1 所示. 孪生神经网络第一个分支的输入图像是模板帧, 以此分支的信息检测候选区域中的目标; 第二个分支的输入图像为检测帧, 即视频后续帧的图像, $\Phi(T)$ 、 $\Phi(S)$ 表示两个分支的输出特征图. 特征图传入 RPN 网络后分别对 $\Phi(T)$ 、 $\Phi(S)$ 再进行一次卷积处理, $\Phi(T)$ 经过卷积后通道数分别提升为原来的 $2k$ 倍和 $4k$ 倍, 而 $\Phi(S)$ 通道数

则保持不变, 随后以 $\Phi(T)$ 为卷积核和 $\Phi(S)$ 做卷积运算, 其输出特征图大小为 17×17 , 以通道数不同分为两个分支: 分类分支和回归分支. RPN 网络在该输出特征图的每个像素点上生成 k 个锚点框, 作为初始框. 在分类分支, 输出特征图包含 $2k$ 个通道, 表示 k 个锚点框的前景与背景的置信度; 在回归分支, 输出特征图包含 $4k$ 个通道, 分别表示 k 个锚点框坐标的偏移量集合. 在回归分支, 线性回归单元以锚点框的坐标参数和偏移量集合为输入对锚点框进行线性回归, 同时引入尺度变化惩罚机制, 根据真实框和每个预测框的差异得到对应的惩罚因子, 惩罚因子对置信度加权, 根据置信度排序提取出最终的预测边框.

2.2 跟踪加速器架构

本文设计的跟踪加速器的系统架构如图 2 所

示,整个系统主要由相机、DMA、软件编译器、片外存储器 DDR、主处理器、神经网络加速器和后处理模块构成。相机用于捕捉视频图像信息,主处理器将采集的信息存储在输入缓存中,卷积核的权重初始化在 DDR 中并通过总线传入权重缓存,编译器同步生成计算指令并传入指令寄存器中。在前向推理阶段,控制器根据指令将输入缓存和权重缓存中的数据分发至神经网络加速器的各个计算阵列,每个

计算阵列由 32 个乘法器和一个加法树构成,共有 32 个计算阵列并行运算,依次完成卷积、池化和激活等操作,随后控制器将神经网络加速器最后一层得到的置信度和参数偏移量的集合传入后处理模块,并行执行不同锚点框的边框回归、尺度惩罚等操作,最后对所有锚点框的置信度排序进行提取,最终的提取结果为预测框的坐标参数 (x_1, y_1, x_2, y_2) ,通过总线传入 DDR 并对视频图像进行标定。

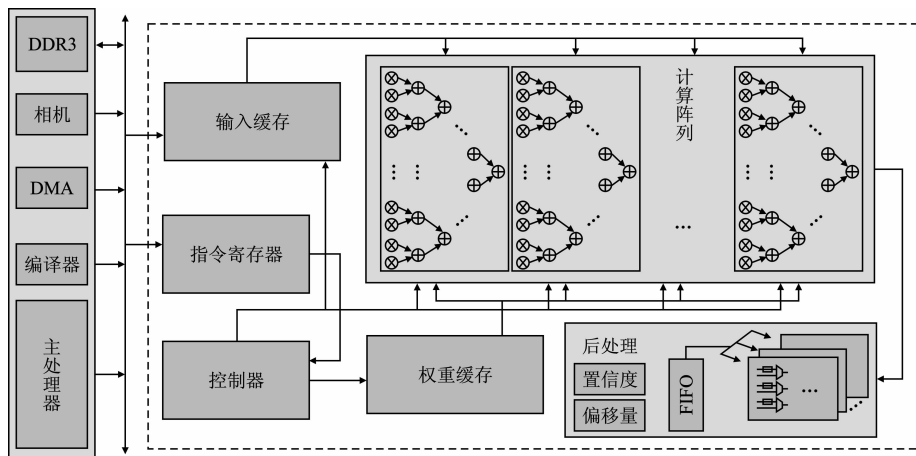


图2 目标跟踪加速器系统架构

3 硬件设计

3.1 神经网络加速器设计

在神经网络加速器中,其计算量主要集中在卷积运算。如图3所示,本节将以1个 $3 \times 3 \times 4$ 大小的输入视图与3个 $2 \times 2 \times 4$ 大小的卷积核为例,详细介绍卷积计算单元中的数据流向及其计算流程,具体如下:首先将输入特征图第一个像素点4个通道的数据分别缓存到4个寄存器中,3个卷积核对应相同位置的权重数据拼接在一起存入行缓存器中,该特征值与行缓存器中的3个权重数据对应相乘,并将3个乘积结果暂存在中间寄存器中;4个通道并行执行上述计算过程,随后将中间寄存器中相同位置的数据传入加法树进行加法运算,并将结果存入输出缓存器中,整个过程耗时3个时钟周期;将上述步骤循环4次,循环执行的顺序先是宽度方向,后是高度方向,12个时钟周期之后输出缓存器中累加得到的数据即为输出特征图的第一个像素点;随后将卷积窗口向右滑动,重复上述过程,24个时钟周期之后卷积核对于第一行输出的计算完成,此时卷积窗口已滑动至最右端;最后在输入特征图上沿高度方向向下滑动卷积窗口,重复上述步骤完成下一行输出的计算。因此仅需要 $24 \times 2 = 48$ 个周期可完

成整张输出特征图的计算。

对于上述的卷积计算单元,无论输入视图和卷积核的尺寸如何,均可支持其卷积计算过程。

3.2 流水线型 Booth 乘法器

硬件实现过程中,乘法器是进行数据处理的关键部件,故乘法器的优化对加速器的性能至关重要。常用的基2 Booth乘法器在进行 n bit数据的乘法运算时会产生 n 个部分积,当操作数的数据位宽越大时,部分积的个数也越多,部分积求和所消耗的逻辑延时将导致运算速度下降。因此本文采用基4 Booth乘法器,它可以将部分积的个数减少一半,有效提高乘法运算的速度。以8 bit数据 M 和 N 的乘法运算为例: $[M * N]_{\text{补}} = M * N_{\text{补}}$,乘数 M 可表示为:

$$M = -m_7 * 2^7 + m_6 * 2^6 + m_5 * 2^5 + m_4 * 2^4 + m_3 * 2^3 + m_2 * 2^2 + m_1 * 2^1 + m_0 * 2^0$$

将上式进行整理可得到如下表达式:

$$M = (-2m_7 + m_6 + m_5) * 2^6 + (-2m_5 + m_4 + m_3) * 2^4 + (-2m_3 + m_2 + m_1) * 2^2 + (-2m_1 + m_0 + m_{-1}) * 2^0$$

其中 $m_{-1} = 0$,由公式可以看出基4 Booth乘法器仅需生成4个部分积,并对4个部分积进行求和即可得到最终的乘积。

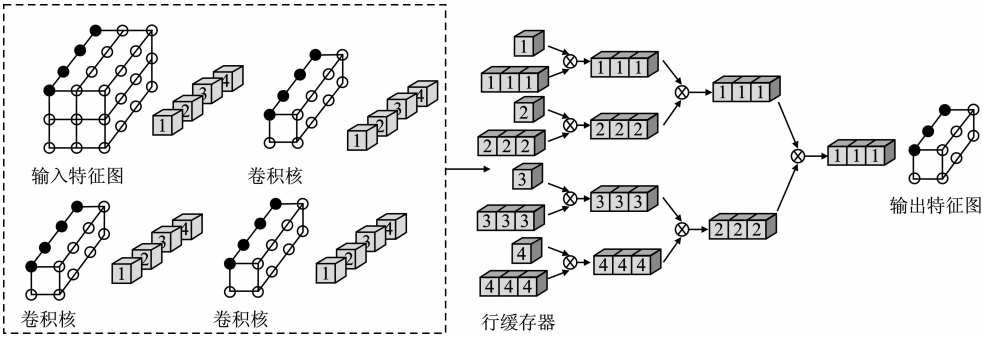


图 3 卷积运算数据流向

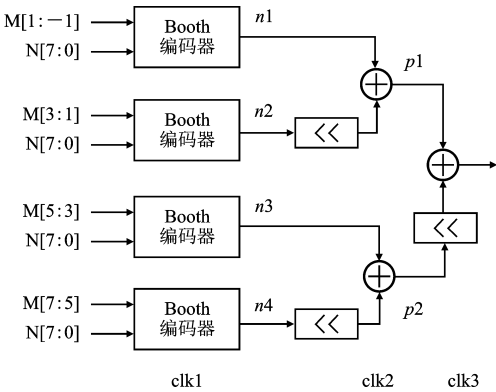


图 4 流水线型基 4 Booth 乘法器

为了进一步提升该乘法器的运算速度,本文引入了流水线型 Booth 乘法器,计算流程如图 4 所示. 该乘法器采用 4 个 Booth 编码器并行计算 4 个部分积: n_1 , n_2 , n_3 和 n_4 ;之后传入加法树对 4 个部分积进行移位求和运算,中间寄存器 p_1 暂存部分积 n_1 与 n_2 的和, p_2 暂存部分积 n_3 与 n_4 的和,随后 p_2 移位与 p_1 求和得到最终的乘积,因此本文设计的 Booth 乘法器仅需要三个时钟周期即可实现 8 bit 数据的乘法运算.

3.3 后处理模块设计

后处理模块是面向目标跟踪任务提出的专用加速模块,其流程图如图 5 所示,对于输出特征图的每一个像素点都对应五个边框需要进行后处理操作,为了节省硬件资源,本文采用分时复用的思想,引入一组硬件串行处理五个边框.

以查找表(Look Up Table, LUT)实现某些复杂非线性运算是常见的硬件优化方式,LUT 是由同一类型的数据元素构成的集合,先把对应函数在固定输入范围内的输出结果写入 LUT,将输入数据量化为查找表地址,实现对应输入的近似计算. 文献 [5]基于查找表设计了近似优化乘法器,乘法运算过程转换为从 LUT 中读取数据的过程,无论是在运算速度还是动态功耗上相较于传统的乘法器都有显著的优势. 由于后处理模块中两次涉及指数函数的求解,该操作对硬件的兼容性具有很大挑战,因此本文选择构建查找表求解单元的方法进行近似求解,这样可以有效减少计算单元求解指数函数所需的时间,实现更好的加速. 根据前向训练结果统计,所涉及的指数函数的输入范围为 0~1 之间,离线生成对应的数据文件,存入查找表以供指数函数运算调用.

4 实验结果与分析

4.1 实验环境

本文采用硬件描述语言编码设计,使用 Synopsys VCS 实现仿真,使用 Synopsys Design Compiler,基于 SMIC 40 nm 工艺对硬件设计进行综合,并根据 PrimeTime PX 仿真得到功耗. 训练环境在 Nvidia K40 上搭建,跟踪数据集为 VOT2015、VOT2016 和 VOT2017,跟踪过程根据准确性和鲁棒性评估性能,使用平均重叠期望评估整体表现.

表 1 不同硬件加速方案的性能对比

	Eyeriss ^[6]	DNA ^[7]	CASSANN-X ^[8]	平台 1	平台 2	本文
工艺技术	ASIC 65nm	ASIC 65 nm	ASIC 28nm	CPU	ASIC 40 nm+RISC	ASIC 40 nm
量化策略	16 bits 定点	16 bits 定点	16 bits 定点	8 bits 定点	8 bits 定点	8 bits 定点
PE 数量	168	512	512	—	1 024	1 024
频率(MHz)	200	200	1 000	2 900	500	500
SRAM	108 KB	280 KB	2.1 MB	—	256 KB	256 KB
功率(W)	0.236	0.479	0.857	157	0.912	0.221
能效比(GOPS/W)	166.20	406.20	582.00	0.70	900.65	5705.73

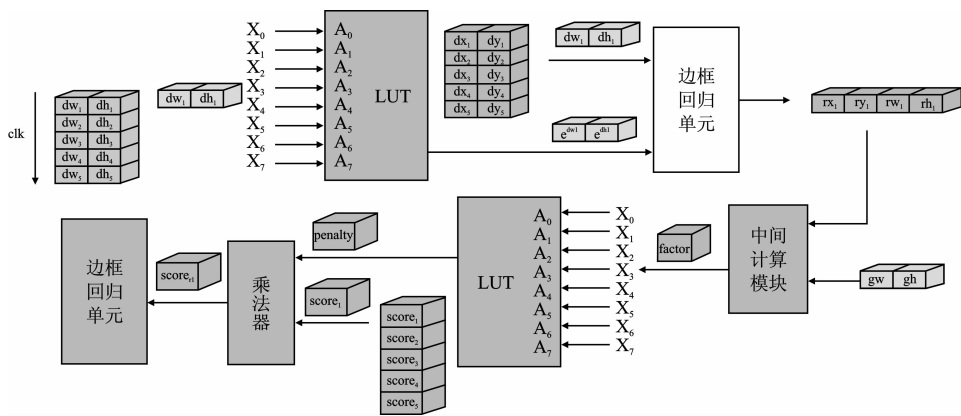


图 5 后处理模块计算流程示意图

4.2 实验分析

为了得到目标跟踪加速器的性能加速比,本文不仅完成了 SiamRPN 算法的硬件评估,而且分别在两种不同的平台上实现了 SiamRPN 算法的跟踪过程.如表 2 所示,平台 1 和平台 2 分别对应 CPU 和神经网络加速器联合 RISC 处理器的实现结果. CPU 处理单帧图像的时间为 368.28 ms,1 秒仅能处理 2.72 帧,无法实现实时处理;神经网络加速器联合 RISC 处理器进行目标跟踪时,单帧图像的处理时间为 49 ms,1 秒可处理 20.41 帧,但仍未实现实时跟踪.由于本设计引入了专用的后处理模块,并联合深度流水神经网络加速器实现数据的端对端处理,相比于其他两种实现方式,本设计的实时处理性更高,每秒可处理 31 帧.与平台 2 相比,本文以其 6.34 倍的能效比获得了 1.53 倍的加速.其中仅针对后处理任务,RISC 处理器的处理时间为 17 ms,本设计提出的专用加速模块以 5.3 ms 的处理时间实现了 3.2 倍的加速,因此该目标跟踪专用架构具有较强的优越性.

目前的芯片多面向分类任务的卷积神经网络进行硬件加速,面向目标跟踪任务的硬件加速设计则比较少,因此本文在进行目标跟踪任务对比的同时,完成了与不同文献神经网络加速器的性能对比,其硬件资源、频率、功率及能效比的对比结果一并在表 1 中列出.本设计的加速器工作频率为 500 MHz,是文献[6]和[7]的 2.5 倍,性能分别为这两种硬件的 58.64 倍和 6.46 倍.文献[8]的工作频率是本设计的 2 倍,但这导致其硬件的反转功率较大,同时较大的片上存储带来了存储访问功率的提升,因此文献[8]的功率最大,约为其他两种硬件的 3.63 倍和 1.79 倍,是本设计的 3.88 倍,因此尽管本设计的频

率不是最高的但其能效比最高,分别为其他三种硬件的 34.33 倍、14.05 倍和 9.80 倍.

4.3 硬件实现

本文设计的目标跟踪加速器的性能参数如表 2 所示,其工作频率为 500 MHz,理论计算峰值达 1.255 TOPS,片上存储空间为 256 KB,其中输入缓存和输出缓存各占 64 KB,权重缓存占 128 KB.采用 GPU 训练网络量化权重后,网络权重和特征像素值均为 8 位定点数,实验表明该量化策略基本不损失跟踪算法的精度.

表 2 芯片性能参数

性能指标参数	参数值
工艺技术	SMIC 40 nm
芯片面积	3.64 mm ²
片上 SRAM	256 KB
PE 数量	1024
时钟频率	500 MHz
精度	8 bits 定点
平均功率	219.79 mW

由综合报告可知,芯片总面积约为 3.64 mm²,平均功耗为 219.79 mW,面积分布示意

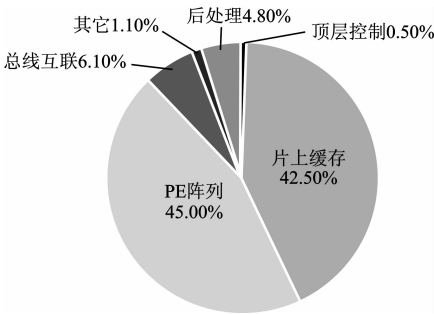


图 6 综合面积分布示意图

图如图 6 所示,主要占比为 PE 阵列和片上缓存,其中 PE 阵列面积占比为 45%,约为 1.64 mm²,片上缓存面积占比为 42.5%,约为 1.55 mm²;后处理模块面积占比仅为 4.8%,约为 0.176 mm²,硬件实现结果表明:本设计在资源消耗较少的情况下,通过引入专用的目标跟踪后处理模块,极大地提高了视频图像的处理速度,以较高的能效比实现了实时处理。

5 结束语

本文深入分析了基于深度学习目标跟踪系统的并行加速原理及可行性,采用流水线型 Booth 乘法器,联合加法树完成卷积运算的加速,结合独立设计的后处理模块分时复用硬件结构实现计算流程,同时通过引入查找表提升计算效率.该硬件加速系统以层级流水线结构实现端到端的数据处理,实现了每秒 31 帧的实时处理,能效比达 5.71 Tops/W,相对于神经网络加速器联合 RISC 处理器实现跟踪任务的方法实现了 1.53 倍的加速.因此本文提出的基于深度学习的目标跟踪加速器适用于相关应用的终端设备。

参考文献:

- [1] HELD D, THRUN S, SAVARESE S. Learning to track at 100 FPS with deep regression networks[C]// European Conference on Computer Vision, Cham: Springer, 2016: 749-765. DOI: 10.1007/978-3-319-46448-0_45.
- [2] 李云松,尹万春,宋长贺.使用 GPU 硬件加速的 Struck 目标跟踪方法:中国,104680558A[P]. 2015-06-03.
[LI Y H, YIN W C, SONG C H. Struck Target Tracking Method Using GPU Hardware for Acceleration: China, 104680558A[P]. 2015-06-03.]
- [3] 季飞,王加庆,刘剑,等.适用于硬件高速计算的 CNN 目标跟踪算法[J]. 微电子学与计算机, 2018, 32(12): 115-118. DOI: 10.19304/j.cnki.issn1000-7180.2018.12.023.
JI F, WANG J Q, LIU J, et al. A target tracking algorithm based on CNN for high-speed calculation of hardware[J]. Microelectronics & Computer, 2018, 32(12): 115-118. DOI: 10.19304/j.cnki.issn1000-7180.2018.12.023.
- [4] LI B, YAN J J, WU W, et al. High performance visual tracking with Siamese region proposal network [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, IEEE, 2018: 1-5. DOI: 10.1109/CVPR.2018.00935.
- [5] 万晨雨,贺雅娟.一种基于存储的乘法器查找表的近似优化方法[J]. 电子产品世界, 2019, 26(7): 44-47.
WAN C Y, HE Y J. Look-up-table approximation for memory-based multiplier design[J]. Outlook of Electronic Technology, 2019, 26(7): 44-47.
- [6] 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. DOI: 10.1109/JSSC.2016.2616357.
- [7] TU F B, YIN S Y, OUYANG P, et al. Deep convolutional neural network architecture with reconfigurable computation patterns[J]. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2017, 25 (8): 2220-2233. DOI: 10.1109/TVLSI.2017.2688340.
- [8] 乔瑞秀,陈刚,龚国良,等.一种高性能可重构深度卷积神经网络加速器[J]. 西安电子科技大学学报, 2019, 46(3): 130-139. DOI: 10.19665/j.issn1001-2400.2019.03.020.
QIAO R X, CHEN G, GONG G L, et al. High performance reconfigurable accelerator for deep convolutional neural networks[J]. Journal of Xidian University, 2019, 46 (3): 130-139. DOI: 10.19665/j.issn1001-2400.2019.03.020.

作者简介:

李 倍 女, (1992-), 硕士研究生. 研究方向为神经网络硬件加速器设计与数字 IC 设计.

闵 丰 男, (1990-), 博士研究生. 研究方向为计算机系统结构与机器人视觉加速器.

杨 军 女, (1995-), 硕士研究生. 研究方向为嵌入式系统及数字 IC 设计.

梁 科 男, (1978-), 博士, 高级实验师. 研究方向为数字集成电路设计.

李国峰(通讯作者) 男, (1958-), 博士, 教授. 研究方向为电路与系统设计及集成电路设计.

E-mail: ligf@nankai.edu.cn