文章编号:1008-1542(2021)03-0241-07

基于 FPGA 的通用卷积层 IP 核设计

安国臣[1],袁宏拓[1],韩秀璐[1],王晓君[1],侯雨佳[2]

(1.河北科技大学信息科学与工程学院,河北石家庄 050018;2.石家庄外国语学校,河北石家庄 050022)

摘 要:针对目前卷积神经网络在小型化、并行化过程中遇到的计算速度不够、可移植性差的问题,根据卷积神经网络和 FPGA 器件的特点,提出了一种利用 VHDL 语言参数化高速通用卷积层 IP 核的设计方法。利用卷积层的计算方式,将卷积核心设计为全并行化、流水线的计算模块,通过在卷积核心的每一行连接 FIFO 的方式改善数据流入的方式,减少地址跳转的操作,并加入控制核心使其可以随图像和卷积窗口大小调整卷积层参数,生成不同的卷积层,最后将卷积层与 AXIS 协议结合并封装成 IP 核。结果表明,在 50 MHz 的工作频率下,使用 2×2 大小的卷积核对 100×100 的图像进行卷积计算,各项资源利用率不超过 1%,耗时 204 μs,计算速度理论上可以达到最高 5 MF/s。因此,设计方案在增加卷积模块可移植性的同时又保证了计算速度,为卷积神经网络在小型化器件上的实现提供了一种可行的方法。

关键词:集成电路技术;卷积神经网络;FPGA;卷积层;设计参数化

中图分类号:TP274;TP391 **文献标识码**:A **doi**:10.7535/hbkd.2021yx03005

Design of universal convolutional layer IP core based on FPGA

AN Guochen[1], YUAN Hongtuo[1], HAN Xiulu[1], WANG Xiaojun[1], HOU Yujia[2]

(1. School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang, Hebei 050018, China; 2. Shijiazhuang Foreign Education Group, Shijiazhuang, Hebei 050022, China)

Abstract: Aiming at the problems of insufficient computing speed and poor portability in the miniaturization and parallelization of convolutional neural network, this paper proposes a design of high-speed universal convolutional layer IP core using VHDL language based on the characteristics of convolutional neural network and FPGA devices. Layer based on convolution calculation, convolution core design is put forward for the parallel calculation and pipeline module, through each line in the convolution of the core connect to FIFO to improve the data flow, reduce the operating address jump, and join the control core to make it can adjust the convolution with images and convolution window size to layer parameters, generate different convolution layer, finally, the convolution layer is combined with the AXIS protocol and encapsulated into IP core. Under the work-

收稿日期:2020-12-29;修回日期:2021-05-07;责任编辑:王淑霞

基金项目:国防科技重点实验室课题(6142205190401)

第一作者简介:安国臣(1977—),男,河北保定人,副教授,硕士,主要从事智能信息处理技术、卫星导航及其关键技术方面的研究。

通讯作者:袁宏拓。E-mail: 781296515@qq.com

安国臣,袁宏拓,韩秀璐,等.基于 FPGA 的通用卷积层 IP 核设计[J].河北科技大学学报,2021,42(3):241-247.

AN Guochen, YUAN Hongtuo, HAN Xiulu, et al.Design of universal convolutional layer IP core based on FPGA[J].Journal of Hebei University of Science and Technology,2021,42(3):241-247.

ing frequency of 50 MHz, the convolution calculation of 100×100 images with 2×2 convolution check is carried out. The utilization rate of each resource is less than 1%, and the time is 204 μs. The theoretical calculation speed can reach the maximum of 5 MF/s. The IP core structure of the convolutional layer not only increases the portability of the convolutional module, but also ensures the computing speed, which provides a feasible implementation method for the implementation of convolutional neural network on miniaturized devices.

Keywords: integrated circuit technology; convolutional neural network; FPGA; convolution layer; design parameterization

随着深度学习研究的发展,卷积神经网络在语音识别[1]、图像理解[2]、目标跟踪[3-4]等领域发挥着重要作用。卷积神经网络既需要在原理和结构上继续研究以达到更好的性能指标和迁移能力,也需要更好地把现有神经网络适用到各种实时性、小型化的场合[5]。越来越多的神经网络加速结构被提出,如将卷积层二值化[6]以减少 FPGA 的运算量、专用的加速 SOC[7]使 CNN 峰值计算性能在 100 MHz 下达到42.13GFLOPS等。基于此提出了一种参数化设计的通用卷积层 IP 核加速结构,该结构适用范围更广。

所提出的通用卷积层 IP 核设计,属于加速器设计通用化,是卷积神经网络加速器设计研究与发展的一个重要方向[8]。该设计通过 VHDL 语言的参数化设计[9],对一种卷积层加速结构[10-11]进行重构,使其由一种定长卷积加速结构改进为可以生成不同大小卷积窗口的通用卷积核。可移植性方面,通过参数配置 IP 核,预先配置其窗口大小,被卷积信号或图像大小,即可生成一维、二维不同大小的卷积层,满足不同的卷积计算需要。计算速度方面,优化数据流入方式,避免了地址的跳转,引入并行化和流水线思想,在数据传输不间断的情况下,一个时钟周期即可输出一个有效计算结果。该设计支持 AXI4-Stream 协议,硬件开发人员可以快速调用该卷积层 IP 核完成卷积层的开发,不必再把精力浪费到内部卷积结构的设计上。

1 卷积层原理

卷积层有一维卷积层、二维卷积层和多维卷积层,同一种卷积层根据卷积窗口大小的不同也不同,这里主要介绍一维卷积层和二维卷积层。

1.1 一维信号卷积

一维卷积通常被用于时间序列的处理,一维卷积神经网络(1-dimensional convolutional neural network,简称 1DCNN)多用于工业故障诊断[12]、医疗诊断[13]等需要对时间序列进行处理的场合。

图 1 为一维卷积示意图,k 个点的时间序列和 n 个点的卷积核做卷积运算,时间序列从左到右滑动,每次滑动对应数据相乘相加输出一个卷积结果,作为下一层卷积层或者池化层的输入。一维卷积运算又等效于 FIR 滤波器的直接型结构。

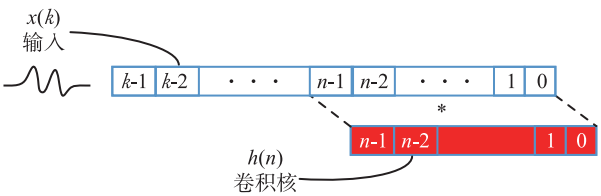

图 1 一维卷积层示意图

Fig. 1 One-dimensional convolutional layer schematic

一维卷积公式如式(1)所示:

$$y(k)=\sum_{i=0}^{n-1}x(i)h(k-1)\ 。\tag{1}$$

式中:$x(i)$ 是输入数据,数据长度为 k;$h(n)$ 是卷积核,卷积长度为 n;$y(k)$ 是卷积核对输入数据进行卷积后的输出,数据长度为 k,在边带不补零的情况下,数据长度为 $k-n+1$。

1.2 二维图像卷积

二维卷积神经网络(2-dimensional convolutional neural network,简称 2DCNN)多用于计算机视觉的处理[14-15],通过卷积核模拟人类大脑的神经元对卷积层输入进行局部感知,模拟人脑神经元感知到生物电信号的反应并将感知结果输出。

图 2 为二维卷积层示意图，一个卷积核模拟的神经元在输入图像上滑动，遍历所有图像数据，每滑动一次，对应的图像数据和卷积核权值相乘求和输出。

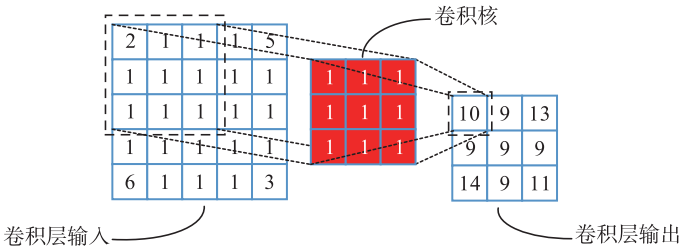

图 2　二维卷积层示意图

Fig. 2　Two-dimensional convolutional layer schematic

二维卷积公式如式(2)所示：

$$y(p,q)=\sum_{j=0}^{m-1}\sum_{i=0}^{n-1}x(i,j)h((p-i),(q-j))\ 。\tag{2}$$

式中：$x(i,j)$ 为输入数据，一般为图像数据；图像大小为 $p\times q$；$h((p-i),(q-j))$ 为卷积核；卷积核的窗口大小为 $m\times n$；$y(p,q)$ 为输出数据，数据长度为 $p\times q$，在边带不补零的情况下，数据长度为 $(p-m+1)\times(q-n+1)$。

2　FPGA 构架分析

从 1988 年提出的 LeNet-5 模型到经典的 VGG-16 模型，卷积层由最初的 2 层增加到了 13 层，甚至在 152 层的 ResNet 网络中达到了 50 层。卷积层数的增加意味着计算量的增加，卷积层的计算量在 CNN 中占比高达 90%[16]。因为神经元和感受野之间的局部连接特点，同一个卷积层下的卷积核是可以并行计算的，同一个卷积核的所有感受区域也是可以并行计算的，所以卷积层可以有很高的并行性。

为了兼顾速度和资源，卷积层一般采用并行加流水线相结合的方式。在某一个感受野内的计算是并行的，即 1 次计算 1 个卷积结果。同一个卷积窗口在输入图像上的滑动是流水线式的，即所有数据依次进入卷积窗口进行计算，既增加了并行度、提高了计算速度，又相对节省资源。

对于一个感受野中的单次卷积运算，以 5×5 大小的卷积窗口为例，对于 FPGA 并行结构的卷积核，在流水线结构下，1 个运算周期可以输出 1 个卷积结果；对于基于冯诺依曼结构或者哈佛结构的通用中央处理器，1 次卷积需要执行 25 次乘法和 24 次加法，共计 49 次计算，加上每次卷积中必须的地址跳转等操作，至少需要 50 个运算周期才能输出 1 个卷积结果。通过对比可知，在相同频率下，5×5 大小的卷积窗口，基于 FPGA 的卷积核的计算速度是基于通用中央处理器的 50 倍以上。

由于传统的 FPGA 电路定制化的特点，针对固定结构设计的卷积核很难移植到其他算法结构中[17-20]，甚至在同一个算法结构中，一个 3×3 的卷积核也很难扩展成 5×5 的卷积核。为了解决卷积层加速结构不便移植的缺点，提出了如图 3 所示的通用卷积层加速结构。由图 3 可知，该卷积层加速结构的通用性在于仅需在通过生成卷积层时，配置几个简单的参数，即可生成含有指定大小卷积窗口的卷积层。3×3，5×5，7×7 甚至是类似于 3×5 非正方形的卷积窗口和 1×N 大小的一维卷积窗口，都可以通过简单的配置生成。

该卷积层加速结构支持 AXI4-Stream 协议，通过 AXI4 总线获取数据。卷积层结构主要分为 3 个部分，分别是控制核心区、数据缓存与预处理区和并行计算区。控制核心区负责接收和产生状态信号，通过 AXI4 总线与外部交互并且通过控制内部 FIFO 的使能信号控制数据的预处理和计算。数据缓存与预处理区在控制信号的控制下，将输入的数据流分别存入不同的 FIFO 中，每一行都配有一个 FIFO。并行计算区不接受控制核心区的控制，在时钟、使能和复位的控制下独立地进行乘累加操作。

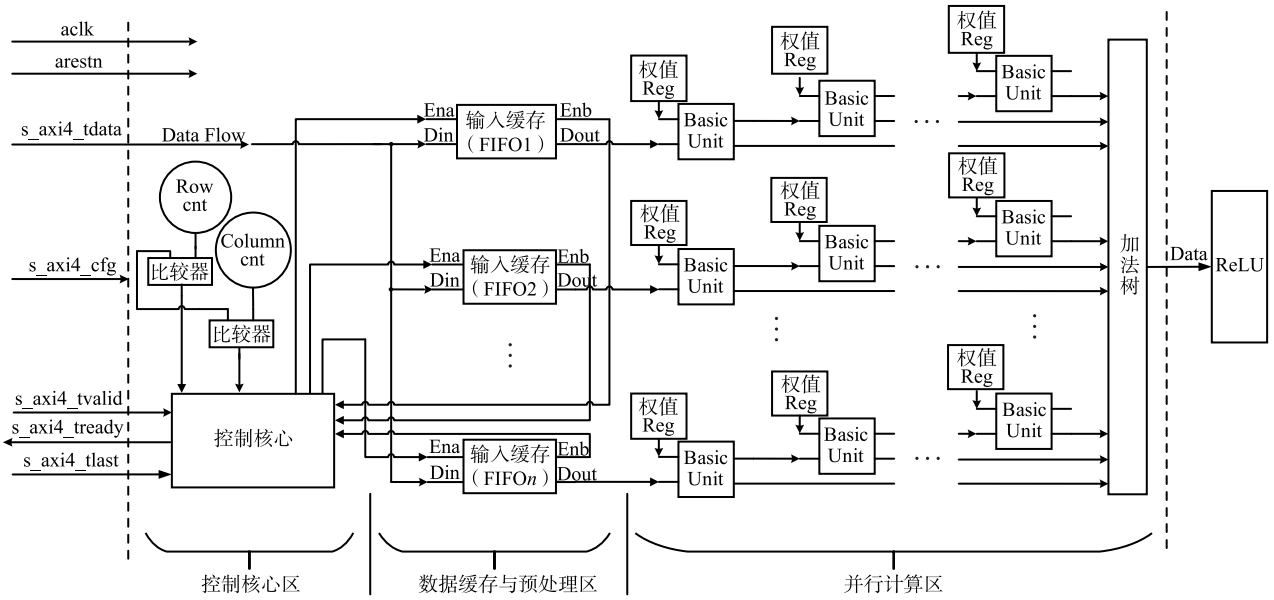

图 3 通用卷积层加速结构

Fig. 3 General convolutional layer acceleration structure

3 通用卷积层 IP 核设计

通用卷积层 IP 核模块运用类属参数语句 generic 和生成语句 generate 相结合的方式进行参数化设计。通过在生成 IP 核时配置参数，将要生成的 IP 核信息传递进入生成语句，生成特定大小的并行计算电路。计算时，将图像的大小信息输入控制核心模块，由控制核心负责切换卷积核的运行状态。

3.1 通用性设计

该卷积核的通用性设计得益于标准的 AXI4-Stream 总线协议和大小可以随参数改变的卷积核结构。AXI4-Stream 总线协议使其可以适配于支持 AXI4-Stream 协议的数据流传递网络中，是为了与 Xilinx 的 IP 核数据交互方式一致。大小随参数改变的卷积核结构使得该卷积核有更高的可重用性，在需要不同窗口大小的网络中生成不同大小的卷积窗口，通用卷积层示意图如图 4 所示。

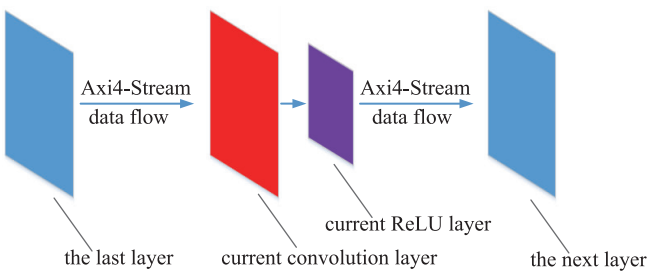

图 4 通用卷积层示意图

Fig. 4 Schematic diagram of the universal convolutional layer

3.2 IP 核设计

该 IP 核的封装设计基于 Xinlinx 公司的 Vivado 2019.1 平台，通过对 IP 核进行封装，可以进一步提高该设计的通用性。通用卷积层 IP 核设计框图如图 5 所示。

该 IP 核由控制核心控制整体时序、Memory 做存储单元、AXIS 到 SDRAM 协议转换的协议转换模块、卷积核心和可选的 ReLU 函数计算模块 5 部分组成。在高速卷积运算核心外围添加 AXIS 协议转换电路，使运算核心支持 AXIS 协议，不仅方便移植，更进一步地，多个支持 AXIS 协议的函数运算单元进行级联可以形成流水线计算模式。比如多个通用卷积层 IP 核模块进行级联，可以形成流水线计算的卷积神经网络。ReLU 函数计算模块作为可选项，在 IP 核配置界面进行配置可以选择是否在该级卷积层后接 ReLU 函数。

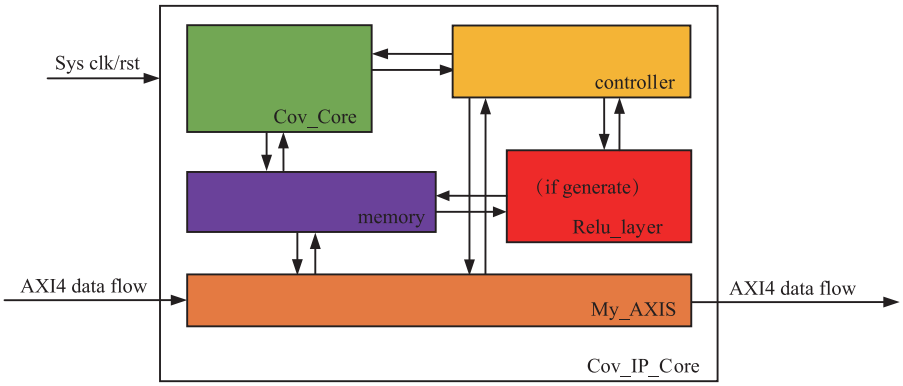

图 5　通用卷积层 IP 核设计框图

Fig. 5　General convolutional layer IP core design block diagram

4　仿真分析

封装后的通用卷积层 IP 核如图 6 所示。左侧为该 IP 核的模组顶层示意图，右侧为该 IP 核的可配置选项。通过配置右侧的参数，包括卷积窗口的大小、数据位宽、输入图像的大小等可以生成特定大小的卷积层。ReLU Mode 选项通过配置 0 和 1，可以选择是否在卷积层后接 ReLU 函数。

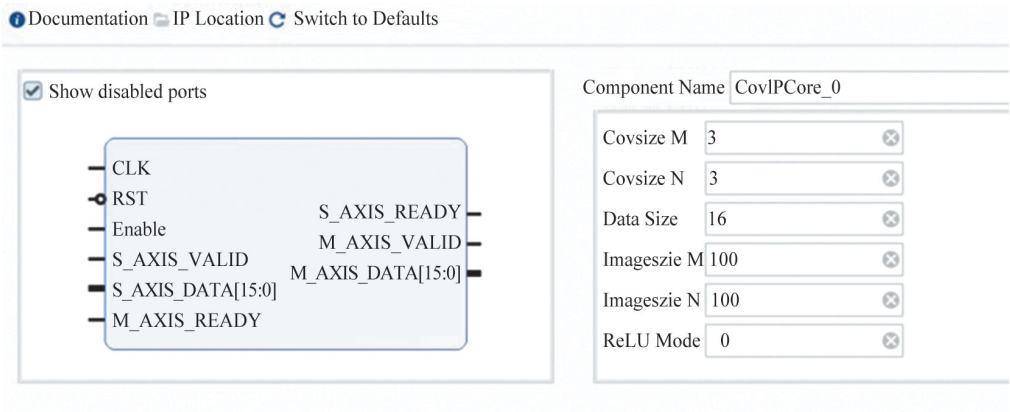

图 6　封装后的通用卷积层 IP 核

Fig. 6　Encapsulated universal convolutional layer IP core

4.1　仿真数据

该设计是在 XILINX 公司的 Vivado 2019.1 开发平台上进行的仿真与分析。模拟生成一幅 100×100 图像的方法验证整个系统的时序逻辑，以下分析都是基于图 5 中 CovCore 模块进行的验证。

模拟生成的源图像有 4 个值，分别是 0x1000，0x2000，0x3000，0x4000，4 个数循环作为激励输入卷积核的输入端，图 7 为图像模拟数据。在仿真中卷积核生成的尺寸为 2×2，按从左到右、从上到下的方向依次滑过源图像。

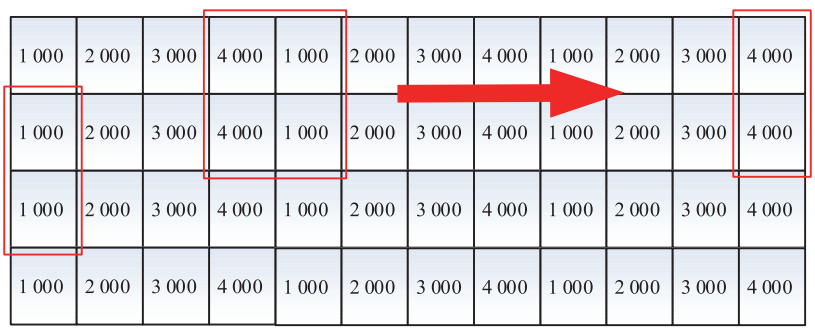

图 7　图像模拟数据

Fig. 7　Image simulation data

卷积核 4 个权值均设为 0x1000，与图像进行卷积后，理论上行首的卷积计算结果应为 0x1800 K，0x2800 K，0x3800 K 和 0x2800 K。图 8 为仿真结果波形图，经对比验证可知电路功能的正确性。

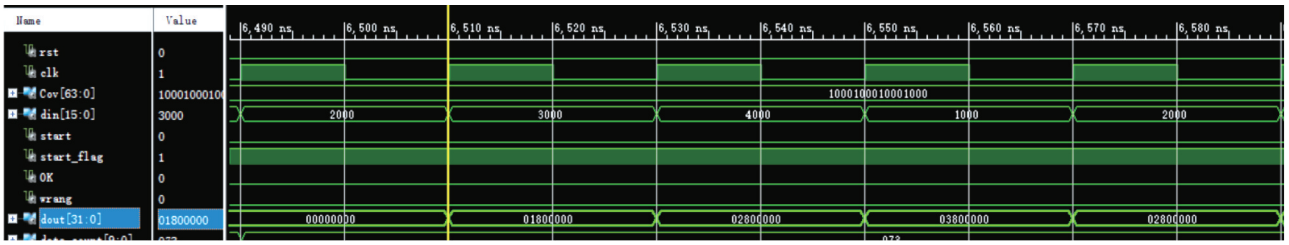

图 8 仿真结果波形图

Fig. 8 Waveform of simulation results

4.2 性能分析

FPGA 芯片型号是 ZYNQ 系列的 xc7z100iffv900-2L，在 Vivado 2019.1 的 Utilization 窗口中查看资源消耗，分别统计了生成 2×2，5×5 和 7×7 大小的卷积核时的资源消耗和计算速度，时钟频率为 50 MHz。表 1 为 FPGA 性能分析。

表 1 FPGA 性能分析

Tab.1 FPGA performance analysis

卷积窗口大小	资源		消耗时间/μs
	LUT	FF	
2×2	40/277 400(0.01)	71/554 800(0.01)	204.120
5×5	66/277 400(0.02)	377/554 800(0.07)	225.180
7×7	72/277 400(0.03)	734/554 800(0.13)	249.200

由表 1 可知，LUT 和 FF 的资源消耗随卷积窗口的增大而增长，其中 FF 资源消耗增长较为剧烈。在计算速度方面，2×2 大小卷积核心的计算时间是 204.120 μs，与 7×7 大小的卷积核心相比速度加快约 45 μs。

在外围电路增加 AXIS 协议转换电路情况下，资源消耗和计算时间又略有增长。LUT 和 FFT 分别增加固定个数 27 和 28，计算时间随着一帧图像数据大小的增长而增长。

5 结 语

在 FPGA 上实现了可变大小的卷积神经网络的卷积层 IP 核，在生成 2×2 大小的卷积层 IP 核时，各项资源利用率不超过 1%，当生成更大的卷积核时，消耗的资源会相应增长。在 50 MHz 的工作频率下，对像素值为 100×100 的图像进行 2×2 卷积窗口的卷积，耗时约 204 μs。生成多级卷积层 IP 核流水线式处理图像数据时，计算速度理论最高可以达到 5 MF/s，即 1 s 处理 5 M 幅 100×100 大小的图像。该通用卷积层 IP 核在满足资源消耗和计算速度的情况下，实现了卷积核的通用性设计，可为卷积神经网络移植和 FPGA 加速构架设计中起到参考作用。

该 IP 核作为专用的卷积层 IP 核和神经网络的基本运算结构，可以被绝大多数深度学习框架所使用。在 AI 芯片、张量处理单元(TPU)等基于深度学习的 ASIC 解决方案中占据重要地位，可以被广泛应用于深度学习加速卡、异构处理器、边缘计算和云计算等当下热门 AI 芯片领域。本文主要针对实数域下的通用卷积层 IP 核结构进行加速设计，未来还需完善将其推广到复数域卷积层 IP 核加速结构，进一步提升通用性。

参考文献/References:

[1] 屈家丽，贺光辉，王国兴.语音识别中卷积神经网络的 FPGA 实现[J].微电子学与计算机，2018，35(9)：37-46.
QU Jiali，HE Guanghui，WANG Guoxing.FPGA implementation of convolutional neural network in speech recognition[J].Microelectronics & Computer，2018，35(9)：37-46.

[2] 常亮，邓小明，周明全，等.图像理解中的卷积神经网络[J].自动化学报，2016，42(9)：1300-1312.
CHANG Liang，DENG Xiaoming，ZHOU Mingquan，et al.Convolutional neural networks in image understanding[J].Acta Automatica

Sinica，2016，42(9)：1300-1312.

[3] 李玺，查宇飞，张天柱，等.深度学习的目标跟踪算法综述[J].中国图象图形学报，2019(12)：2057-2080.
LI Xi，ZHA Yufei，ZHANG Tianzhu，et al.A review of deep learning target tracking algorithms[J].Journal of Image and Graphics，2019(12)：2057-2080.

[4] PEI Xia，LI Dong，WANG Lijun，et al.Deep visual tracking：Review and experimental comparison[J].Pattern Recognition.2018，76：323-338.

[5] 李彦冬，郝宗波，雷航.卷积神经网络研究综述[J].自动化学报，2016，36(9)：2508-2515.
LI Yandong，HAO Zongbo，LEI Hang.A review of convolutional neural networks[J].Acta Automatica Sinica，2016，36(9)：2508-2515.

[6] 蒋佩卿，吴丽君.基于 FPGA 的改进二值化卷积层设计[J].电气开关，2019(6)：8-13.
JIANG Peiqing，WU Lijun.Design of improved binarization convolutional layer based on FPGA[J].Electric Switchgear，2019(6)：8-13.

[7] 赵烁，范军，何虎.基于 FPGA 的 CNN 加速 SoC 系统设计[J].计算机工程与设计，2020，41(4)：939-944.
ZHAO Shuo，FAN Jun，HE Hu.Design of CNN acceleration SoC system based on FPGA[J].Computer Engineering and Design，2020，41(4)：939-944.

[8] 陈桂林，马胜，郭阳.硬件加速神经网络综述[J].计算机研究与发展，2018，56(2)：240-253.
CHEN Guilin，MA Sheng，GUO Yang.A review of hardware-accelerated neural networks[J].Journal of Computer Research and Development，2018，56(2)：240-253.

[9] 孙延腾，吴艳霞，顾国昌.基于 VHDL 语言的参数化设计方法[J].计算机工程与应用，2010，46(31)：68-71.
SUN Yanteng，WU Yanxia，GU Guochang.Parametric design method based on VHDL language[J].Computer Engineering and Applications，2010，46(31)：68-71.

[10] 刘志成，祝永新，汪辉，等.基于 FPGA 的卷积神经网络并行加速结构设计[J].微电子学与计算机，2018.35(10)：80-84.
LIU Zhicheng，ZHU Yongxin，WANG Hui，et al.Design of convolutional neural network parallel acceleration structure based on FPGA[J].Microelectronics & Computer，2018，35(10)：80-84.

[11] 陈煌，祝永新，田犁，等.基于 FPGA 的卷积神经网络卷积层并行加速结构设计[J].微电子学与计算机，2018，35(10)：85-88.
CHEN Huang，ZHU Yongxin，TIAN Li，et al.Design of convolutional layer parallel acceleration structure of convolutional neural network based on FPGA[J].Microelectronics & Computer，2018.35(10)：85-88.

[12] 安晶，艾萍，徐森，等.一种基于一维卷积神经网络的旋转机械智能故障诊断方法[J].南京大学学报(自然科学)，2019(1)：133-142.
AN Jing，AI Ping，XU Sen，et al.An intelligent fault diagnosis method for rotating machinery based on one-dimensional convolutional neural network[J].Journal of Nanjing University(Natural Science)，2019(1)：133-142.

[13] 黄佼，宾光宇，吴水才.基于一维卷积神经网络的患者特异性心拍分类方法研究[J].中国医疗设备，2018(3)：11-14.
HUANG Jiao，BIN Guangyu，WU Shuicai.Patient - specific cardiopap classification based on one-dimensional convolutional neural network[J].China Medical Devices，2018(3)：11-14.

[14] 王礼贺，杨德振，李江勇，等.卷积神经网络在目标检测中的应用及 FPGA 实现[J].激光与红外，2020，50(2)：252-256.
WANG Lihe，YANG Dezhen，LI Jiangyong，et al.Application of convolutional neural network in target detection and FPGA implementation[J].Laser and Infrared，2020，50(2)：252-256.

[15] 江泽涛，刘小艳，胡硕.基于 CNN 的红外与可见光融合图像的场景识别[J].计算机工程与设计，2019，40(8)：2289-2294.
JIANG Zetao，LIU Xiaoyan，HU Shuo.Scene recognition of infrared and visible fusion images based on CNN[J].Computer Engineering and Design，2019，40(8)：2289-2294.

[16] HE K，ZHANG X，REN S，et al.Deep residual learning for image recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).New York：IEEE，2016：770-778.

[17] 王婷，陈斌岳，张福海.基于 FPGA 的卷积神经网络并行加速器设计[J].电子技术应用，2021，47(2)：81-84.
WANG Ting，CHEN Binyue，ZHANG Fuhai.Design of convolutional neural network parallel accelerator based on FPGA[J].Application of Electronic Technique，2021，47(2)：81-84.

[18] 张旭欣，张嘉，李新增，等.二值 VGG 卷积神经网络加速器优化设计[J].电子技术应用，2021，47(2)：20-23.
ZHANG Xuxin，ZHANG Jia，LI Xinzeng，et al.Accelerator optimization design of binary VGG convolutional neural network[J].Application of Electronic Technique，2021，47(2)：20-23.

[19] 张帆.图像卷积实时计算的 FPGA 实现[J].电子设计工程，2021，29(1)：132-137.
ZHANG Fan.FPGA implementation of image convolution real-time computation[J].International Electronic Elements，2021，29(1)：132-137.

[20] 范军，巩杰，吴茜凤，等.基于 FPGA 的 RNN 加速 SoC 设计与实现[J].微电子学与计算机，2020，37(11)：1-5.
FAN Jun，GONG Jie，WU Xifeng，et al.Design and implementation of RNN accelerated SoC based on FPGA[J].Microelectronics & Computer，2020，37(11)：1-5.