基于硬件仿真系统的边缘计算人工智能视觉芯片设计验证

徐宣哲[1,2]，宁珂[1,2]，郑学敏[1,2]，赵明心[1,2]，徐萌萌[1,2]，吴南健[1,2]，刘力源[1,2]

（1. 中国科学院半导体研究所半导体超晶格国家重点实验室，北京 100083；
2. 中国科学院大学材料与光电研究中心，北京 100049）

摘 要：基于卷积神经网络（CNN, convolutional neural network）的视觉深度学习算法的兴起推动了人工智能视觉芯片设计研究的快速发展，而芯片的设计验证工作是人工智能视觉芯片研发的瓶颈。介绍了一种基于硬件仿真系统的人工智能视觉芯片软硬件验证方法，以边缘计算人工智能视觉芯片设计为例，在硬件仿真系统 ZeBu 上完成了芯片运行的典型深度学习网络 MobileNet 的仿真验证工作。结果表明，在硬件芯片架构上实现的网络模型在保证精确度的同时，在 200 MHz 频率时钟下单帧检测时间只需要 18.51 ms，与软件平台仿真相比，仿真速度提高了 7 倍。

关键词：人工智能视觉芯片；深度学习；MobileNet；ZeBu
中图分类号：TN47
文献标志码：A
DOI:10.11959/j.issn.2096−3750.2022.00250

Verification of an artificial intelligence vision chip design for edge computing based on hardware simulation system

XU Xuanzhe[1,2], NING Ke[1,2], ZHENG Xuemin[1,2], ZHAO Mingxin[1,2], XU Mengmeng[1,2], WU Nanjian[1,2], LIU Liyuan[1,2]

1. The State Key Laboratory of Superlattices and Microstructures, Institute of Semiconductors, Chinese Academy of Sciences, Beijing 100083, China;
2. Center of Materials Science and Optoelectronics Engineering, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: The rise of visual deep learning algorithms based on convolutional neural network (CNN) has promoted the rapid development of the artificial intelligence (AI) vision chip design research. The step of chip verification is a bottleneck in the development of AI vision chips. A software and hardware verification method for AI vision chip design based on hardware simulation system was introduced. Taking AI vision chip design for edge computing as an example, the chip was run on the hardware simulation system (ZeBu) and the simulation verification work of typical deep learning network MobileNet was completed. The results show that the network model implemented on the hardware chip architecture keeps accuracy while the detection time of a single frame is only 18.51 ms under a 200 MHz clock frequency. The spread of the hardware simulation is 7 times faster than than of the software simulation.

Key words: AI vision chip, deep learning, MobileNet, ZeBu

0 引言

卷积神经网络（CNN, convolutional neural network）AlexNet[1]被提出以来，其在图像处理领域的潜力被研究人员不断地挖掘[2]。得益于 CNN 具有优异的集成度、适应性和特征提取能力，目前基

收稿日期：2021−05−19；**修回日期：**2022−03−01
通信作者：刘力源，liuly@semi.ac.cn
基金项目：国家重点研发计划（No.2019YFB2204300）；国家自然科学基金资助项目（No.U20A20205，No.61874107）；中国科学院青年创新促进会项目（No.2021109）
Foundation Items: The National Key Research and Development Program of China (No.2019YFB2204300), The National Natural Science Foundation of China (No.U20A20205, No.61874107), The Program of Youth Innovation Promotion Association, Chinese Academy of Sciences (No.2021109)

于该类结构的多种网络已经成为了图像识别[3]、目标检测[4-7]、目标追踪[8]、语音识别[9]等信号处理工作的主流选择。

随着诸如 SCA-CNN[10]、ShuffleNet[11]等更高精度的图像处理算法的出现，卷积神经网络的可应用面变得越来越广泛[12]。然而，神经网络模型的日趋复杂化在带来准确率红利的同时，对于搭载设备运算能力的要求也越来越高——一个完整的卷积神经网络需要高性能服务器运行数天乃至更长时间进行参数优化的工作，优化完成后对测试集和实际图像的处理即使在高性能的图形处理器（GPU, graphics processing unit）上运行也需要花费较长时间，这无疑制约了卷积神经网络在算力相对不强的硬件端的发展前途。

基于硬件的小规模内存和小规模处理单元，研究人员设计了一批专门针对硬件特点提出的卷积神经网络[13]。该类神经网络具有参数量小、卷积过程运算量小的特点，其中又以 MobileNet 系列[14]最为典型。同时，为了适应高速发展的高效率神经网络，开发了一些传感—存储—运算一体化，人工智能视觉芯片架构，如中国科学院半导体研究所开发的高速可编程视觉芯片[15-16]等，它能够在进行视觉图像传感的同时完成智能化的视觉信息处理。

即使在小型芯片上，实际设计过程中仍然需要很长时间来完成芯片的设计和验证工作，其中验证工作又占据了超过七成的时间——完整的验证过程需要对片上基本的指令、神经网络的操作、具体算法和应用场景进行串行验证，将每个片上实现的算法逐一验证是非常漫长的过程。EDA 公司提供的软件验证方式在 CPU 上运行，但是在 CPU 上运行仿真过于缓慢，周期过长，因此另一种更加快速的验证方式——在硬件仿真系统 ZeBu[17]上运行——成为了更优的选择。

基于上文所述的趋势，本文首先对硬件仿真系统 ZeBu 及其上的验证流程进行说明，然后介绍了一种边缘计算人工智能视觉芯片架构和针对此架构设计进行轻量化后的 MobileNet V1 结构及其在硬件上操作的实现过程，最后以此架构和网络为例对于其 ZeBu 上的验证工作进行了详细的阐述。本文将着重介绍 MobileNet V1 算法的轻量化操作、硬件实现和在 ZeBu 系统进行的硬件仿真器级验证的相关工作。

本文的主要创新点体现在以下 3 个方面。

1) 提出了一种模型大小仅为 2.9 MB 的超轻量级分类卷积神经网络算法；

2) 提出了一种针对该分类算法的高效利用片上资源的硬件代码实现方法；

3) 使用 ZeBu 进行硬件仿真器仿真，更真实和高效地完成了验证工作。

1　ZeBu 系统的使用方法和验证流程

ZeBu 系统是一个硬件层次的仿真验证平台，由 PC 端服务器和 ZeBu 服务器两个部分构成。ZeBu 服务器上的资源实现了硬件互联级别的硬件代码映射，而 PC 端服务器则通过外部设备互联（PCI, peripheral component interconnect）端口对 ZeBu 服务器生成的硬件模型进行控制。

ZeBu 系统相较于传统的现场可编程门阵列（field-programmable gate array）FPGA 测试板验证而言具有更高的集成度：后者需要在与 FPGA 相连的上位机中编写输入数据和控制信号，而 ZeBu 平台则将这一工作集成到了平台内部，只需要将运行软件寄存器转换级（RTL, register transfer level）仿真时的配置文件输入，交由 ZeBu 系统的 PC 端进行分别处理操控输入即可。同时，ZeBu 系统采用了互动组合信号运算（iCSA, interactive combined signal arithmetic）技术，该技术使得从载入到开始发生输入信号相较于传统仿真波形发生器从几个小时缩减到了几分钟；其运行后调试的模式使得对逻辑分析仪跟踪窗口的限制解除，测试者可在不对设计进行重新编译的情况下再次运行测试[18]。

在 ZeBu 系统上的验证流程简图如图 1 所示。首先，PC 上位机将视觉芯片硬件代码下载至 ZeBu 系统，在系统中完成视觉芯片模型的烧写工作，然后将算法网络结构数据和需要用到的测试数据加载入视觉芯片模型，在模型上运行算法进行测试，测试的结果将会按序输出至上位机。将输出结果与理想结果进行比对即可验证算法和芯片模型的正确性，如果不一致则可以通过 ZeBu 配套的工具查看 ZeBu 系统的外围信号波形来进行调试。

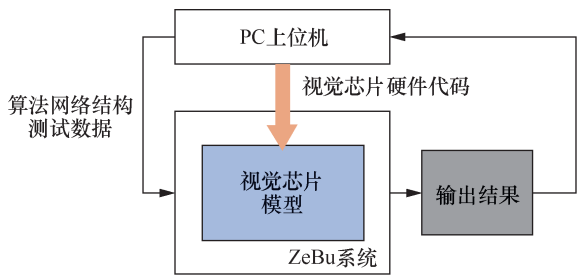


图 1　在 ZeBu 系统上的验证流程简图

2 验证边缘计算人工智能视觉芯片的框架

边缘计算人工智能视觉芯片模型如图 2 所示，其核心部分为微控制器（MCU, micro control unit）、视觉处理器（VP, vision processor）及存储器（VPdm, vision processor data memory）。该芯片的输入为移动产业处理器接口（MIPI, mobile industry processor interface）和四分串行外设接口（QSPI, quarter serial peripheral interface），输出则是 MIPI 和两线式串行总线（I2C, inter-integrated circuit）master 端。输入的 MIPI 负责接收来自图像传感器的数据，QSPI 负责从片外的 flash memory 之中读取信息；输出的 I2C master 端负责输出对图像传感器的控制信号，输出的 MIPI 则负责将 VP 处理完成后的数据向片外输出。此芯片设计的寄存器转换及电路（RTL，register transfer level）代码会由 ZeBu 系统烧写，芯片运行的时钟频率为 200 MHz。

一次完整的运行流程为：首先，所有的数据被输入到片外 flash memory 之中，然后 MCU 控制代码和 VP 控制代码，被预设程序自 QSPI 读入至对应模块中开始运行；MCU 将图像传感器的配置和控制信号自 I2C 接口发送给图像传感器，并控制着图像传感器进行成像操作，后将图像数据自 MIPI 送回至处理部分，经过前端根据 MCU 中参数控制的信号处理后，数据存入 VPdm；得到图像信息后，MCU 控制 VP 算法参数也从片外 flash memory 读入 VPdm 中。所需要的数据都已被写入 VPdm 后，MCU 控制 VP 开始运行，VP 处理完成后，MCU 从 VPdm 中存储运行结果的地址读取对应长度的数据自 MIPI 输出至片外。

仿真系统并不对图像传感器的行为进行仿真，因此图像输入来自于单独的图像文件，从 MIPI 读入。

3 验证的算法：轻量化的 MobileNet V1

3.1 MobileNet V1 结构简述

MobileNet V1 是一种轻量级的卷积神经网络，其结构建立在深度可分离卷积（DSC, depthwise separable convolution）之上，是由普通 4-D 卷积层（4-D convolution）、深度卷积层（depthwise convolution）、点式卷积层（pointwise convolution），平均池化层（average pooling）[19]和全连接层（full connect）[20]组成的共 29 层的网络结构，其中 28 层为卷积层。MobileNet 各层特征图（feature map）尺寸和卷积核尺寸见表 1（以输入图像的尺寸为 224 px×256 px×3 px 为例）。

从表 1 可以看出，MobileNet 在最终输出 1 024 类时的参数大小仅为 16 MB，这得益于深度可分离卷积对参数量的有效减小，使用一层深度卷积层和一层点式卷积层来近似实现标准 4-D 卷积的效果。以第 14～15 层为例计算数据量如下（该式计算的是参数量，每个参数占用 4 byte，因此实际占用大小需要乘以 4）：

$$\mathrm{Size}_{\mathrm{normal}} = \mathrm{Length}^2_{\mathrm{kernel}} \cdot \mathrm{Dim}_{\mathrm{input}} \cdot \mathrm{Dim}_{\mathrm{output}} = 3\times 3\times 512\times 512 = 2\,304\ \mathrm{KB}$$

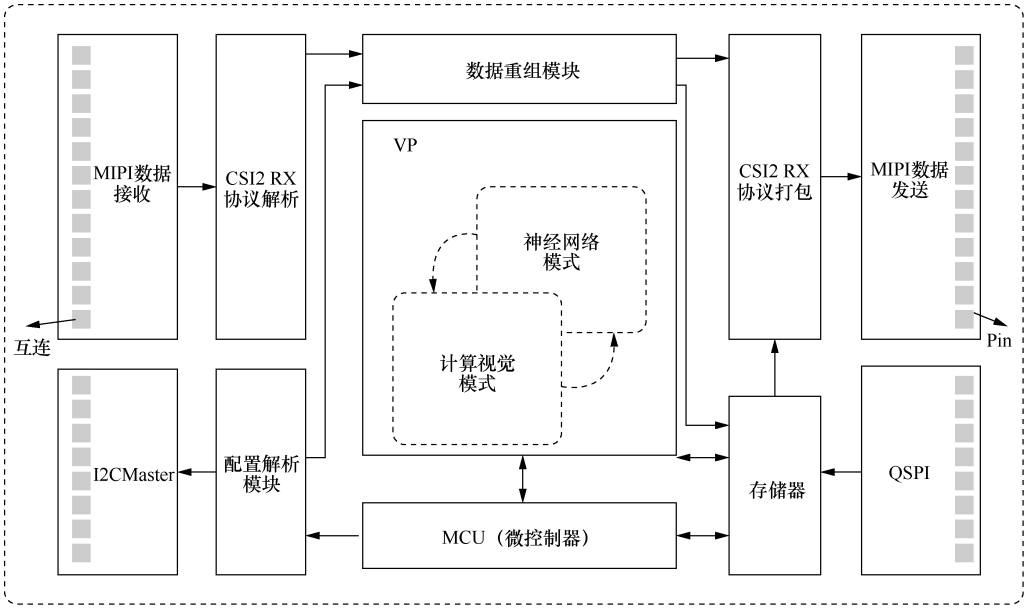


图 2 边缘计算人工智能视觉芯片模型

表 1　　MobileNet 各层特征图（feature map）尺寸和卷积核尺寸

层数	类型	输出特征图 大小（尺寸）	卷积核 大小（尺寸）
1	4-D 卷积层	0.438 MB（112px×128 px×32 px）	3.375 KB（3×3×3×32）
2	深度卷积层	0.438 MB（112 px×128 px×32 px）	1.125 KB（3×3×32）
3	点式卷积层	0.875 MB（112 px×128 px×64 px）	8 KB（1×1×32×64）
4	深度卷积层	0.219MB（56 px×64 px×64 px）	2.25 KB（3×3×64）
5	点式卷积层	0.438 MB（56 px×64 px×128 px）	32 KB（1×1×64×128）
6	深度卷积层	0.438 MB（56 px×64 px×128 px）	4.5 KB（3×3×128）
7	点式卷积层	0.438 MB（56 px×64 px×128 px）	64 KB（1×1×128×128）
8	深度卷积层	0.109 MB（28 px×32 px×128 px）	4.5 kB（3×3×128）
9	点式卷积层	0.219 MB（28 px×32 px×256 px）	128 KB（1×1×128×256）
10	深度卷积层	0.219 MB（28 px×32 px×256 px）	9 KB（3×3×256）
11	点式卷积层	0.219 MB（28 px×32 px×256 px）	256 KB（1×1×256×256）
12	深度卷积层	0.055 MB（14 px×16 px×256 px）	9 KB（3×3×256）
13	点式卷积层	0.109 MB（14 px×16 px×512 px）	512 KB（1×1×256×512）
14～23	深度卷积层	0.109 MB（14 px×16 px×512 px）	18 KB（3×3×512）
	点式卷积层	0.109 MB（14 px×16 px×512 px）	1 024 KB（1×1×512×512）
24	深度卷积层	0.027 MB（7 px×8 px×512 px）	18 KB（3×3×512）
25	点式卷积层	0.055 MB（7 px×8 px×1 024 px）	2 048 KB（1×1×512×1024）
26	深度卷积层	0.055 MB（7 px×8 px×1 024 px）	36 KB（3×3×1024）
27	点式卷积层	0.055 MB（7 px×8 px×1 024 px）	4 096 KB（1×1×1024×1024）
28	平均池化层	0.001 MB（1 px×1 px×1 024 px）	-
29	全连接层	0.001 MB（1 px×1 px×1 024 px）	4 096 KB（1×1×1024×1024）

$$\mathrm{Size}_{\mathrm{depthwise}} = \mathrm{Length}_{\mathrm{kernel}}^{2} \cdot \mathrm{Dim}_{\mathrm{input}} \cdot 1 = 3 \times 3 \times 512 \times 1 = 4.5\ \mathrm{KB}$$

$$\mathrm{Size}_{\mathrm{pointwise}} = 1^{2} \cdot \mathrm{Dim}_{\mathrm{input}} \cdot \mathrm{Dim}_{\mathrm{output}} = 1 \times 1 \times 512 \times 512 = 256\ \mathrm{KB}$$

可见使用深度卷积层和点式卷积层替换标准卷积可以缩减约 8/9 的参数数据量，同时最大的单层数据量也显著下降，这使得该算法更加适用于硬件上的实现——对硬件上的存储空间要求下降，设计硬件芯片时可以优先考虑其他方面的性能。

3.2　MobileNet V1 轻量化处理

尽管 MobileNet 原始版本的数据量在深度可分离卷积神经网络中已经属于相对偏低的数值，但加上特征图所需要的空间后超过 17 MB 的总空间需求量依旧超出了很多硬件端高速存储单元大小的限制，同时过大的数据体量对于硬件端口有限的传输速度也是一个挑战，会导致读入参数的进程过长，影响硬件的表现性能。

由于本文设计的芯片是一款基于边缘应用的视觉芯片，其存储空间受限，因此从系统需求的角度出发，为了高效地部署在硬件上，本文在设计算法时对 MobileNet V1 进行了轻量化裁剪。

轻量化处理主要通过两个操作实现：一是对 MobileNet V1 进行减层操作，去除表一中 14～23 层循环操作中的两组深度卷积层和点式卷积层，在保证精度最大限度不受损失的情况下实现对参数大小的缩减；二是对该网络进行 8 bit 量化工作，将网络中的参数数据精度从 32 bit 削减为 8 bit。

经过两项网络轻量化操作后，改造后的 MobileNet V1 总参数量被缩减为 2.9 MB，可以与相邻两层的输出输入特征图同时置于总空间为 4.5 MB 的存储器之中。轻量化后的网络比原始网络在大小上缩减了 82.2%，精度的损失则仅为 1.96%，在

Caltech-101 数据集下的精确度保持在 83.33%。部分知名网络的参数量对比见表 2。

表 2　部分知名网络的参数量对比

序号	网络名称	参数量/MB
1	AlexNet[1]	60
2	ZFNet[21]	62.3
3	VGG16[22]	138.3
4	VGG19[22]	143.6
5	Inception V3[23]	23.8
6	ResNet50[24]	25.5
7	MobileNet V1[14]	16
8	NASNetMobile[25]	5
9	轻量化 MobileNet V1（本文）	2.9

4　MobileNet V1 在边缘计算人工智能视觉芯片上的实现

轻量化操作后的 MobileNet V1 输入图像要求为 RGB 逐通道排列的范围为−128～127（8 位有符号数）的二维矩阵，对应尺寸为 224 px×256 px×3 px；而经过如图 2 所示的数据重组模块的处理后，从传感器输入的图像由拜尔阵列格式的二维矩阵变为了按 GBR 顺序逐行排列的 0～255 的二维矩阵，其对应图像尺寸为 548 px×960 px×3 px。因此，在进行 MobileNet V1 运算之前需要进行图像转换操作实现图像大小和数值的匹配。下面将对图像转换操作和神经网络卷积操作的实现分别进行说明，两个操作的算法设计都保证了最大的并行度，设计时将所有可并行的操作并行处理，并以最高的 VP 存储单元占用率进行计算，以此保证算法的高效性。

4.1　图像转换操作的实现

为了实现图像数据的转换，本文在计算视觉模式下的视觉处理器上完成了两项操作：第一项操作为映射，将按 GBR 顺序逐行排列的二维矩阵对应的图像尺寸从 548 px×960 px×3 px 映射为 224 px×256 px×3 px；第二项操作为图像重排，将图像从按 GBR 顺序逐行排列，转换为按 RGB 顺序逐通道排列，并将数据范围进行转换。

映射操作的硬件实现流程如图 3 所示，该操作主要通过 5 个步骤实现，每项操作均基于 8 个为一组的寄存器进行设计以保证高效完成。

图像重排操作的硬件实现流程如图 4 所示，该操作主要分为两步，先进行像素数据排列转换，再进行数据大小变换。

4.2　神经网络卷积操作的实现

在卷积神经网络中，各层主要进行的计算为卷积，具体到 MobileNet V1，各层主要进行的计算为标准 4-D 卷积，深度卷积和点式卷积 3 种模式。本节将对于这 3 种卷积的硬件实现进行简述。

各卷积的循环示意图如图 5 所示，卷积实现（1-4-D 卷积，2-深度卷积，3-点式卷积）如算法 1 所示，标准 4-D 卷积由 3 层嵌套的循环完成：在最里层循环（循环 1）中，每次读入输入特征图（ifmap, input feature map）一个维度中 3 行的数据和该输入输出特征图维度对应的卷积核进行卷积，得到一行指定维度输出特征图（ofmap, output feature map）中来自该 ifmap 维度的分量；在第 2 层循环中，循环 1 的输入维度部分进行了累加并经过了偏置、乘子和移位操作，得到了指定 ofmap 维度中一行的数据；在第 3 层循环中，循环 2 的数据进行了反复读

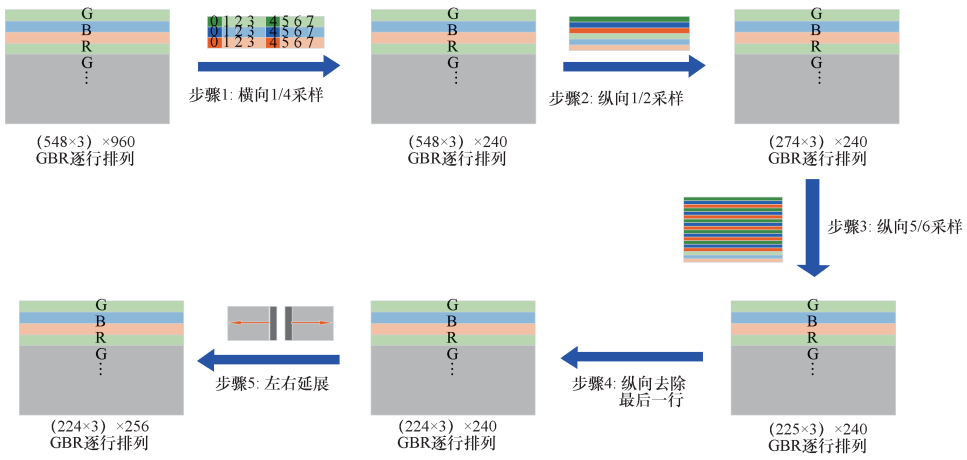


图 3　映射操作的硬件实现流程

出的过程，最终输出了指定 ofmap 维度的数据；循环 3 循环完毕时，所有维度的 ofmap 输出组成了整体的 ofmap 结果输出。

算法 1　卷积实现（1-4-D 卷积，2-深度卷积，3-点式卷积）

Define: VP Regfiles Data o

1) Repeat-loop3

2)　Repeat-loop2

3)　Load (Bias_Data)

4)　Load (Mul_Data)

5)　Load (Shift_Data)

6)　Repeat-loop1

7)　Load(Weight_Data)

8)　Load(Ifmap_Data)

9-1)　o ← o+Convolution (Weight_ Data,Ifmap_Data)

9-2)　o ← o+Convolution (Weight_ Data, Ifmap_Data)

9-3)　o ← o+Weight_Data* Ifmap_Data

10-1)　Until all ifmap dimensions have been calculated

10-2)　EndRepeat（Only calculate once）

10-3)　Until all ifmap dimensions have been calculated

11)　o ← Shift(o*Mul_Data)

12)　Save(o)

13)　initialize(o)

14)　Until all lines of that dim ofmap have been output

15) Until all ofmap dimensions have been output

深度卷积与标准 4-D 卷积的区别在于，深度卷积 ofmap 的指定维度数据只来自于 ifmap 相同维度（M=N）的数据和对应的参数数据，因此，这个改动映射到硬件操作中体现为最里层循环无须进行累加，原来的第 1 层循环卷积结果就是 ofmap 对应维度的一行数据，而第 2 层和第 3 层循环的设置与之前保持一致。

点式卷积和标准 4-D 卷积的区别在于，点式卷积的卷积核大小为 1×1，因此在最里层循环中无须进行标准卷积操作，只需使用乘法即可完成同样的效果，其余结构关系与标准 4-D 卷积保持一致。

5　MobileNet V1 的硬件验证过程和结果

ZeBu 系统的仿真流程如图 6 所示。在硬件方面，代表芯片设计的 Verilog 代码被输入 ZeBu 的编译器进行规划，并将规划好的命令流输入 ZeBu 服

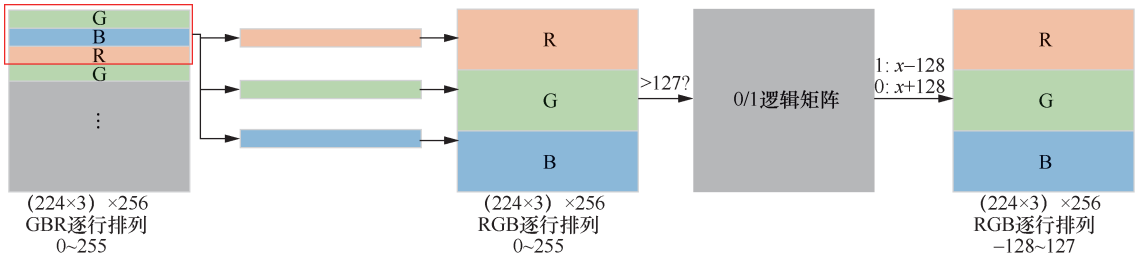
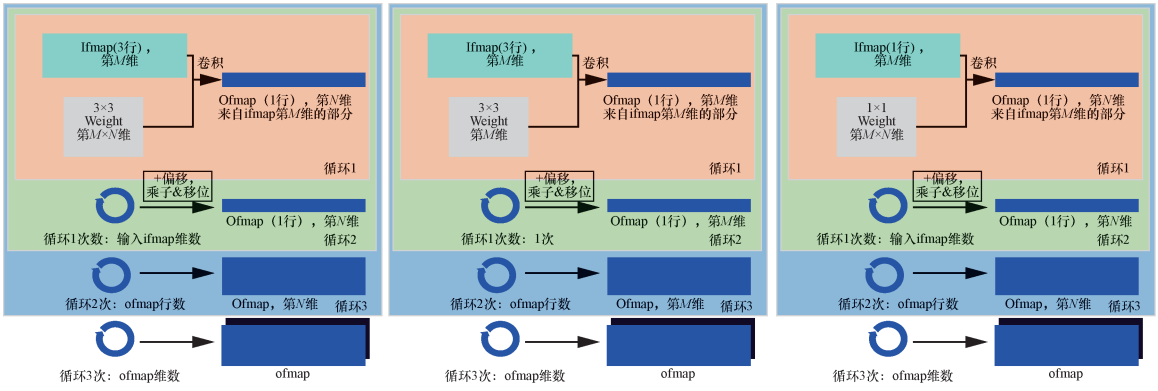


图 4　图像重排操作的硬件实现流程



（a）标准4-D卷积　　（b）深度卷积　　（c）点式卷积

图 5　各卷积的循环示意图

务器进行实际硬件的烧写工作，生成芯片模拟单元。在算法仿真方面， MCU 代码、VP 代码和算法参数的模拟 flash memory 数据和代表图像传感器输入的图像数据被输入 ZeBu 的 PC 服务器，在其上模拟 flash memory 和图像传感器输出数据的行为与芯片模拟单元进行联合仿真，最终将得到的模拟芯片运算结果作为平台输出提供给用户。在本文所述的验证过程中，ZeBu 的工作时钟频率设定为 3 225 kHz，驱动时钟数量为 1 个。

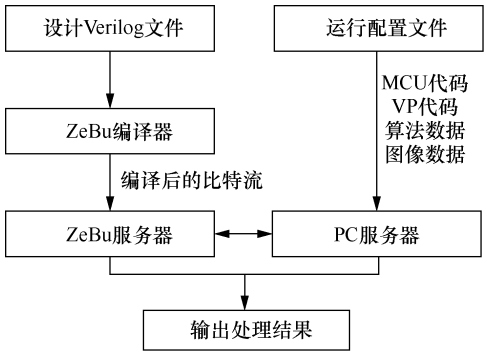


图 6　ZeBu 系统的仿真流程

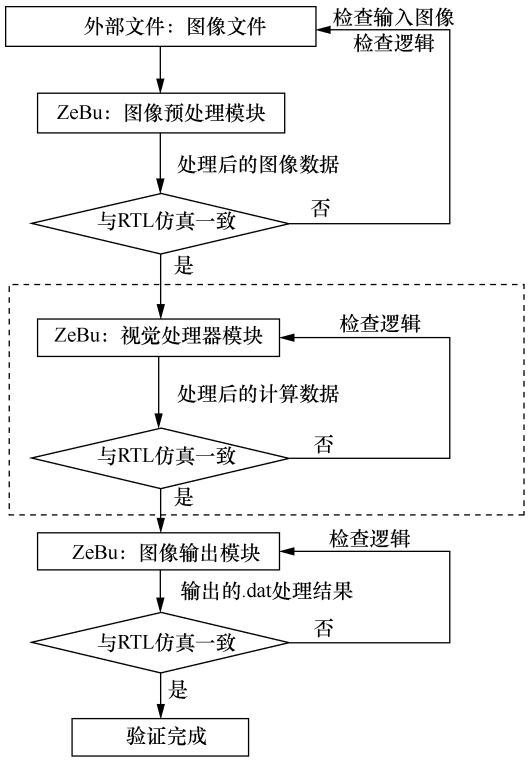


图 7　硬件算法在 ZeBu 系统的验证流程

硬件算法在 ZeBu 系统的验证流程如图 7 所示。图像预处理模块主要涉及输入图像数据、图像处理参数和 RTL 代码。输入图像数据的对比可以通过软件端的数据对比完成，而图像处理参数和 RTL 代码是输入到硬件之中进行处理的，因此需要在 ZeBu 系统中使用逻辑信号查看软件 Verdi 进行各项数据的确认和调试工作。

各项在 VP 上进行的硬件算法模块操作和处理完成后的图像输出模块操作的验证工作同样需要在软件端实现，并对软硬件的输出进行对比。这部分操作主要涉及芯片运行逻辑的验证，当对比结果不一致时，需要逐一对 MCU 的指令、VP 的指令和 RTL 代码进行检查。同时，由于 ZeBu 系统是在实际硬件中由理想时序转为实际时序进行运行的，因此可能出现时钟抖动等引发的时序问题。

本文采用 Caltech-101 数据集进行验证，共 1 020 张（102 类，每类 10 张）测试集图像。部分测试集图像分类结果如图 8 所示。轻量化后的 MobileNet V1 分类准确率为 83.33%。软件端和 ZeBu 端输出和中间结果的逐一对比显示，各平台的分类算法数据一致，这说明二者实现了同样的功能，验证通过。

各硬件算法的 EDA 工具和 ZeBu 运行时间见表 3，得益于仅有 2.9 MB 的轻量化算法设计，MobileNet V1 算法需要运行 3 702 028 个时钟周期，以 200 MHz 时钟频率为基准换算得到单幅片上运行时间约为 18.51 ms，每秒可检测的图像数为 54 幅。从表 3 可以看出，ZeBu 端的仿真运行时间远低于 EDA 软件端，二者运行时间相差 7 倍左右，ZeBu 作为芯片验证工具具有强大的时间优势；同时，ZeBu 可以直接将 EDA 软件端的输入作为其输入，无须重新建立测试文件，作为硬件仿真

图 8　部分测试集图像分类结果

器级仿真工具，相较于传统的 FPGA 板上仿真，在流程和时间上具有优越性。

表 3　各硬件算法的 EDA 工具和 ZeBu 运行时间

算法类型	单幅片上运行时钟周期数	单幅EDA运行时间/min	单幅 ZeBu 运行时间/min
图像分类	3 702 028	114.20	15.73
人脸检测	12 241 338	377.65	51.85
目标追踪	4 996 608	152.85	21.23
TOF 成像	4 560 400	140.65	19.33

6　结束语

本文首先介绍了硬件仿真系统 ZeBu 的简要结构和验证操作流程，然后，介绍了一种边缘计算人工智能视觉芯片架构和轻量化后的 MobileNet V1 结构及其在硬件上操作的实现过程，最后，以此架构和网络为例，对其 ZeBu 上的验证工作进行了详细的阐述。本文的主要亮点在于超轻量级分类神经网络算法设计、高效的硬件代码实现以及使用 ZeBu 加速硬件仿真器级仿真过程。

本文的验证结果表明，在硬件算法上成功复现了本文所提出的轻量化 MobileNet V1 模型，其单帧检测时间可达 18.51 ms，在更小资源的硬件平台上实现了 MobileNet V1 算法并在集成度更高、搭建工程更简易的硬件仿真器级验证工具 ZeBu 上完成了验证工作。如今，越来越多的神经网络算法被应用于小资源硬件端，验证步骤的优化也是其中非常重要的研究课题之一，下一步工作将会继续优化各类轻量级神经网络模型并使用 ZeBu 工具完成更多的仿真。

参考文献：

[1] KRIZHEVSK Y A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communication softhe ACM, 2017, 60(6):84-90.

[2] SZEGED Y C, LIU W, JIA Y Q, et al. Going deeper with convolutions[C]//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 1-9.

[3] DONG X D, XU Y P, XU Z J, et al. A static hand gesture recognition model based on the improved centroid watershed algorithm and a dual-channel CNN[C]//Proceedings of 2018 24th International Conference on Automation and Computing (ICAC). Piscataway: IEEEPress, 2018: 1-6.

[4] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[5] SHIN H C, ROTH H R, GAO M C, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning[J]. IEEE Transactions on Medical Imaging, 2016, 35(5): 1285-1298.

[6] LIN T Y, ROYCHOWDHURY A, MAJI S. Bilinear CNN models for fine-grained visual recognition[C]//Proceedings of 2015 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2015: 1449-1457.

[7] WE IH, ZHU M, WANG B, et al. Two-level progressive attention convolutional network for fine-grained image recognition[J]. IEEE Access, 2020 (8): 104985-104995.

[8] ZHANG Y, YANG S Y, LI H B, et al. Shadow tracking of moving target based on CNN for video SAR system[C]//Proceedings of IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium. Piscataway: IEEE Press, 2018: 4399-4402.

[9] WANG M, ABDELFATTAH S, MOUSTAFA N, et al. Deep Gaussian mixture-hidden Markov model for classification of EEG signals[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2018, 2(4): 278-287.

[10] CHEN L, ZHANG H W, XIAO J, et al. SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 6298-6306.

[11] ZHANG X Y, ZHOU X Y, LIN M X, et al. ShuffleNet: an extremely efficient convolutional neural network for mobile devices[C]// Proceedings of 2018 IEEE/CVFConference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 6848-6856.

[12] SUDHA S, JAYANTHI K B, RAJASEKARAN C, et al. Segmentation of RoI in medical images using CNN-A comparative study[C]//Proceedings of TENCON 2019 - 2019 IEEE Region 10 Conference. Piscataway: IEEE Press, 2019: 767-771.

[13] SHARMA A K, FOROOSH H. Slim-CNN: alight-weight CNN for face attribute prediction[C]//Proceedings of 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition. Piscataway: IEEE Press, 2020: 329-335.

[14] HOWARD A G, ZHU M L, CHEN B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications[EB]. 2017.

[15] SHI C, YANG J, HAN Y, et al. A 1000 fps vision chip based on a dynamically reconfigurable hybrid architecture comprising a PE array processor and self-organizing map neural network[J]. IEEE Journal of Solid-State Circuits, 2014, 49(9): 2067-2082.

[16] LIHL, ZHANGZ X, YANGJ,et al. A novel vision chip architecture for image recognition based on convolutional neural network[C]// Proceedings of 2015 IEEE 11th International Conference on ASIC. Piscataway: IEEE Press, 2015: 1-4.

[17] VINAY B K, HARIHARM, KILLEDAR A. The FPGA based emulation of complex SoC for ADAS market on ZeBu-Server[C]//Proceedings of 2014 International Conference on Advances in Electronics Computers and Communications. Piscataway: IEEE Press, 2014: 1-4.

[18] ENDC. Synopsys 推出业界最快的仿真系统[EB]. 2014.
ENDC. Synopsys Iaunches the industry’s fastest simulation system[EB]. 2014.

[19] WANG S H, JIANG Y Y, HOUXX, et al. Cerebral micro-bleed detection based on the convolution neural network with rank based average pooling[J]. IEEE Access, 2017 (5): 16576-16583.

[20] BASHA S H S, DUBEY S R, PULABAIGARI V, et al. Impact of fully connected layers on performance of convolutional neural networks for image classification[J]. Neurocomputing, 2020, 378: 112-119.

[21] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks[C]//Computer Vision – ECCV 2014, 2014: 818-833.

[22] SIMONYANK, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB]. 2014.

[23] SZEGEDY C, VANHOUCKEV, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 2818-2826.

[24] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 770-778.

[25] ZOPH B, VASUDEVAN V, SHLENS J, et al. Learning transferable architectures for scalable image recognition[C]//Proceedings of 2018 IEEE/CVFConference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 8697-8710.

[作者简介]

徐宣哲（1999– ），男，中国科学院半导体研究所半导体超晶格国家重点实验室硕士生，主要研究方向为低功耗视觉芯片的硬件算法验证等。

宁珂（1994– ），男，中国科学院半导体研究所半导体超晶格国家重点实验室博士生，主要研究方向为低功耗视觉芯片设计等。

郑学敏（1995– ），男，中国科学院半导体研究所半导体超晶格国家重点实验室博士生，主要研究方向为低功耗视觉芯片设计等。

赵明心（1992– ），男，中国科学院半导体研究所半导体超晶格国家重点实验室博士生，主要研究方向为卷积神经网络的量化、稀疏化、加速以及软硬件协同设计等。

徐萌萌（1997– ），女，中国科学院半导体研究所半导体超晶格国家重点实验室博士生，主要研究方向为边缘计算型视觉芯片的网络模型压缩等。

吴南健（1961– ），男，博士，中国科学院半导体研究所半导体超晶格国家重点实验室研究员，主要研究方向为大规模数模混合集成电路和视觉芯片等。

刘力源（1982– ），男，博士，中国科学院半导体研究所半导体超晶格国家重点实验室研究员，主要研究方向为高速图像传感器、太赫兹图像传感器、视觉芯片以及数模混合信号集成电路设计等。