多分支卷积神经网络的 FPGA 设计与优化

谢思璞，魏榕山
（福州大学 物理与信息工程学院，福建 福州 350108）

摘 要：针对拓宽神经网络的结构会导致计算量增大，计算性能降低，需要针对并行的网络进行更有效的优化以及调度。通过分析 FPGA 平台上实现卷积神经网络的计算吞吐量和所需的带宽，在计算资源和访存带宽的限制下，采用了屋顶模型进行了设计空间的探索，提出了在不同支的卷积神经网络中使用不同的循环展开因子，从而实现同一卷积层中不同支神经网络的并行计算，保证计算资源和内存资源的合理分配。实验结果表明，所提出的设计与先前研究相比获得了 1.31×的性能提升。

关键词：多分支卷积神经网络；FPGA；屋顶模型；并行计算
中图分类号：TN409　　**文献标识码**：A　　DOI：10.16157/j.issn.0258-7998.211279

中文引用格式：谢思璞，魏榕山. 多分支卷积神经网络的 FPGA 设计与优化[J].电子技术应用，2021，47(7)：97-101.
英文引用格式：Xie Sipu，Wei Rongshan. FPGA design and optimization of multi-branch CNN[J]. Application of Electronic Technique，2021，47(7)：97-101.

FPGA design and optimization of multi-branch CNN

Xie Sipu，Wei Rongshan
(School of Physics and Information Engineering，Fuzhou University，Fuzhou 350108，China)

Abstract：Broadening the structure of the neural network will lead to the increase of the amount of computation and the decrease of the computational performance，it is necessary to optimize and schedule the parallel network more effectively. By analyzing the throughput and bandwidth of convolutional neural network on FPGA platform，the roof model is used to explore the design space under the limitation of computing resources and memory access bandwidth. It is proposed to use different cycle expansion factors in different branches of convolutional neural network，so as to realize the parallel computing of different branches of neural network in the same convolution layer and ensure the computing efficiency reasonable allocation of resources and memory resources. The experimental results show that the performance of the proposed design is improved by 1.31×compared with the previous research.

Key words：multi-branch convolutional neural network；FPGA；roofline model；parallel computing

0 引言

近年来，神经网络受到了广泛热议，成为了学术界和工业界的热门议题，Google、Microsoft 和 Facebook 等科技公司都建立了相关的研究小组，以探索 CNN 的新架构[1-3]。通过对 CNN 架构上的创新改善 CNN 性能，利用空间和通道信息，结构的深度和宽度以及多路径信息处理等方法引起了广泛的讨论。

在众多新型 CNN 架构中，基于宽度扩展的多支并行的 CNN 得到了国内外学术届的重视。KAWAGUCHI K 等人提出网络的宽度是影响网络精度与准确度的一个重要指标[4]。通过在层中并行使用多个处理单元，可以得到比感知器更为复杂的映射。GoogLeNet 中的 Inception 模块是一种典型的多支网络架构，并使用了不同尺寸的卷积核[5]。2017 年，DEL COCO M 等人[6]利用多分支结构引入了并行的多尺度分析，减小了神经网络的深度，克服了过拟合问题。拓宽网络宽度的多支并行卷积神经网络在图像分割以及识别等任务中，提高了网络在不同尺度上的特征提取能力，受到了国内外研究机构的重视[7-9]。

如今，基于 FPGA 的卷积神经网络加速器获得越来越多的关注[10-13]，目前大多数都以切片的方式，映射至加速器中逐一计算。如果使用相同配置的加速器对多分支网络进行加速，传统映射多分支网络计算会造成硬件资源的浪费。没有针对各分支网络进行差异化资源分配，多分支卷积神经网络在进行时的并行效率也无法得到保证，这对整体网络模型计算将造成影响。本文分析和探讨了多分支卷积神经网络的性质和特点，并基于屋顶模型进行设计空间的探索[14]，设计出一种多分支卷积神经网络加速器。

1 多分支 CNN 模型及参数

本文将以应用于 HEVC 视频编码的帧内划分的多分支卷积神经网络 ETH-CNN 为例，分析与探究多支卷积神经网络在参数差异影响下的硬件资源优化问题[15]。

如图 1 所示，该网络由三支卷积神经网络协同工作，各分支的卷积神经网络经过降采样层、卷积层和全连接层后输出。

由于各分支卷积层的输入特征图尺寸不同，那么各分支网络卷积层的计算量也不同。以 ETH-CNN 的各分支网络参数(如表 1 所示)进行比较，经计算后的结果可以看出各分支的操作数相差甚大，各分支网络在进行并行计算所需的 I/O 带宽以及硬件资源的分配需求不一样。使用相同的加速器对多分支网络进行加速，各卷积层将依次调用加速器进行计算，由于支与支之间卷积存在数据依赖，这将对整体网络模型计算造成影响。为了提升整体算力，为多分支并行卷积神经网络专门定制一种加速平台提供高性能计算是非常有必要的。

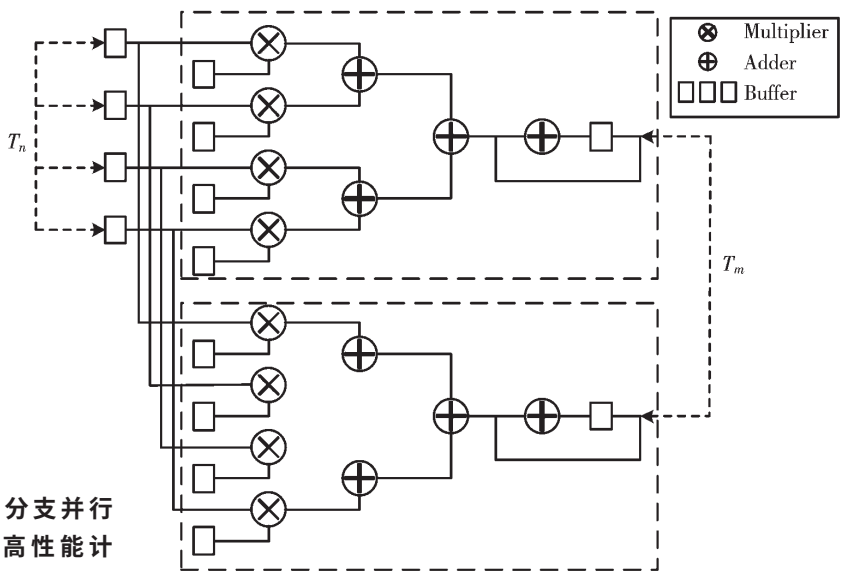

图 2　加速器计算单元

2 多分支 CNN 加速器设计

2.1 CNN 加速器设计

本文将选择在卷积核的输入输出尺度上进行循环拆分，拆分的因子为 T_m、T_n。外部的循环则负责调度，主要的作用是控制数据从外部 DDR 加载至 FPGA 的 BRAM 缓冲区，将 BRAM 数据写入外部 DDR，同时控制数据的复用。内部的循环则负责生成 FPGA 卷积计算单元。设计的卷积计算单元的并行度为 $T_m \times T_n$，并行 T_m 计算单元，每个计算单元中进行 T_n 输入数据与权重的乘累加操作，累加操作中采用加法树结构进行累加。在本设计中利用卷积计算在输入通道和输出通道之间的独立性，采用流水线操作，进一步加大系统的吞吐量。

为了实现高并行度，神经网络加速器通常在大量计算单元之间重用数据。在本设计中，将数据路由到不同的计算单元，从而降低系统计算延迟，为此采取了“前-后-前”的调度模式，从而使输入数据可以获得更大的数据复用机会。原先输入数据读取遵循切分先后，调整后的访问的次数将比原先少 (T_n-1) 次，从而可以减小数据的冗余访问。在输出复用的调度中，由于输出的结果数组与输入通道循环没有相互依赖性，因此可以将写入外部 DDR 的输出结果操作置于输入通道循环外。此时，计算单元无需先从外部 DDR 读取上次的输出结果。输出结果的中间值直接存储在片上的 BRAM 上，本次计算无需上一次的计算结果，直接从 BRAM 上读取数据进行累加，从而可以将输出数据的访问次数减少。

在本设计中，还将采用双缓冲机制，两块片上的缓存以乒乓机制进行数据的读取，从而可以让数据的传输

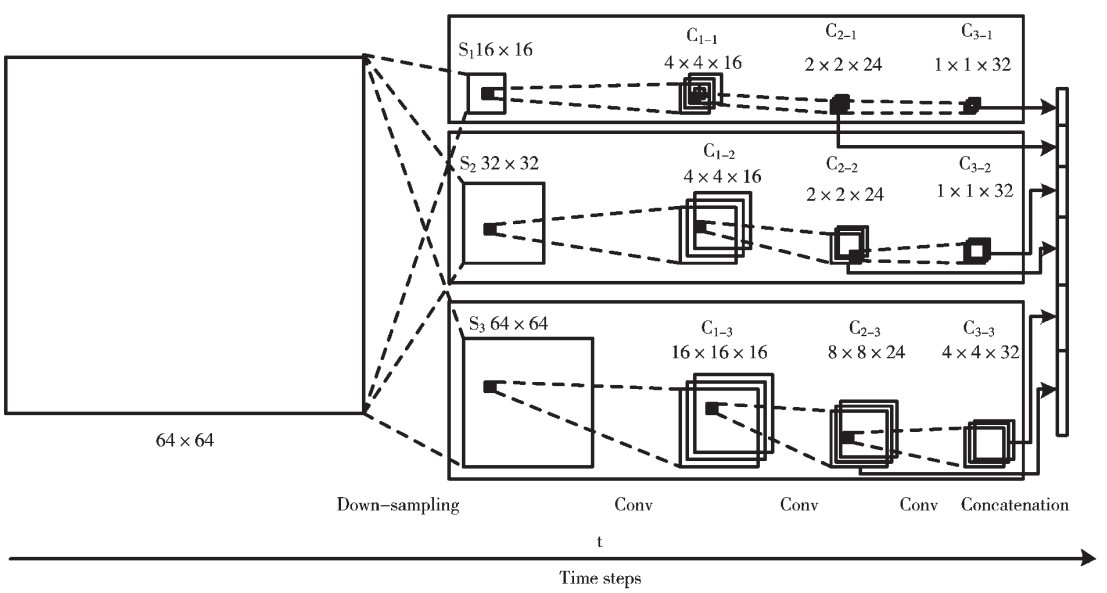

图 1　ETH-CNN 图示

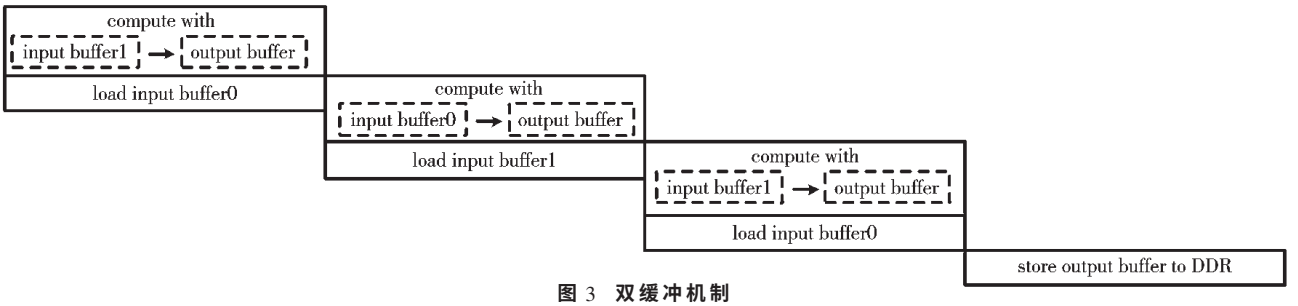

图 3　双缓冲机制

时间被计算单元的计算时间所覆盖，使计算单元一直处在工作状态，提高系统的吞吐率。

2.2 设计空间探索

由于多支卷积神经网络计算是同时进行的，计算和存储访问行为将变得更加复杂。在加速器的设计中，可以利用屋顶模型来解决多支网络模型和硬件平台相关的优化问题。为了并行化循环的执行，将循环展开，在硬件上并行化计算单元。硬件上并行化的参数称为 unroll 参数。展开参数选择不当可能会导致严重的硬件未充分利用。以单循环为例，假设循环的循环边界为 M，并行度为 m。硬件的利用率受$(M/m)/\lceil M/m \rceil$的限制。

多支卷积结构已在先前给出。本文为三支 CNN 都设计了并行度不同的计算单元。在进行一层计算时，三个计算单元同时并行执行，而对层与层之间的计算，则共用同一加速器，即加速器中含有三个并行的加速单元，分别对应三支的 CNN，而这个加速器是层与层之间的共用模块。根据上文设计的卷积计算单元，分别对应三支 CNN 输入输出通道的拆分因子$\langle T_m, T_n \rangle$进行讨论。从整个系统层面上来看，每次卷积的执行周期与分支网络中执行周期最大的一支密切相关。

加速器的峰值性能的计算公式如式(1)所示。其中 total operations 为多分支卷积神经网络计算总操作数，execution cycles 为加速器的执行周期，而在多分支 CNN 加速器中，卷积层的总执行周期与分支网络中执行周期最大的一支密切相关。CTC Ratio 是单位内存访问可以执行的操作数，total data access 为片上计算的外部数据访问量。在多分支 CNN 加速器中，可以通过之前介绍的数据复用设计，降低对片外访存的次数，从而提高系统的 CTC Ratio。

$$\text{Attainable Performance}=\frac{\text{total operations}}{\text{execution cycles}} \tag{1}$$

$$\text{Computation to Communication Ratio}=\frac{\text{total operations}}{\text{total data access}} \tag{2}$$

确定了多分支网络的并行策略后，要针对各项因素设立约束条件。由于加速器的计算峰值受到 FPGA 片上 DSP48E 计算资源的限制，因此其设计空间的限制条件为式(3)。此外，加速器的 CTC Ratio 受到 FPGA 片上BRAM 的限制，所以其存储空间的限制条件为式(4)。

$$\begin{cases} 0<\sum_{i=1}^{b}\text{DSP}_b \leqslant \text{DSP}_{\text{total}},\ \forall\, b\in(0,\text{branches}] \\ 1\leqslant T_{n_b}\leqslant N_b,\ \forall\, b\in(0,\text{branches}] \\ 1\leqslant T_{m_b}\leqslant M_b,\ \forall\, b\in(0,\text{branches}] \end{cases} \tag{3}$$

$$\begin{cases} 0<\sum_{i=1}^{b}\text{BRAM}_b \leqslant \text{BRAM}_{\text{total}} \\ \sum_{i=1}^{b}\text{BRAM}_b=\sum_{i=1}^{b}(B_{\text{in}_b}+B_{\text{wgt}_b}+B_{\text{out}_b}) \end{cases} \tag{4}$$

此外，也需要对循环展开的利用率进行约束，避免系统运行时出现过多的硬件资源处于空闲状态导致了硬件平台的资源浪费，对系统执行效率造成负面影响。为此，对各分支网络的循环展开利用率进行阈值约束，设 α 为循环展开利用率，为其设定循环展开利用率阈值，舍去利用率值低的循环展开因子。

$$\alpha_b=\frac{\frac{M_b}{T_{m_b}}\times\frac{N_b}{T_{n_b}}}{\left\lceil\frac{M_b}{T_{m_b}}\right\rceil\times\left\lceil\frac{N_b}{T_{n_b}}\right\rceil},\ \forall\, b\in(0,\text{branches}] \tag{5}$$

而在多分支卷积神经网络执行过程中，虽然整体系统的执行周期以各分支中执行周期最长的卷积神经网络为基准。但是，为了协调各分支卷积神经网络层与层之间的计算与数据的交换，对分支网络各卷积层的执行时间进行约束。根据式(6)可知，可以通过各分支计算操作数与展开因子作为评估条件，控制各分支网络中最快执行周期与最慢执行周期的卷积神经网络差值，保证整体执行效率处于最优效果。

$$\begin{cases} \text{ExeTime}_{b_i}\propto\frac{M_{b_i}\times N_{b_i}\times R_{b_i}\times C_{b_i}\times K_{b_i}\times K_{b_i}}{T_{m_b}\times T_{n_b}} \\ \forall\, b\in(0,\text{branches}] \\ \forall\, i\in(0,\text{layers}] \end{cases} \tag{6}$$

通过以上提出的峰值和 CTC 率的计算方法，可以在屋顶模型中对不同的拆分因子$\langle T_m, T_n \rangle$对系统造成的影响进行量化以及比较，采用枚举算法可以直观地在屋顶模型中看到设计的性能优劣。图 4 中的每一个点都代表三组拆分因子$\langle T_m, T_n \rangle$，在 FPGA 有限的资源和设计空间限制下对所有的合法的设计进行了枚举，从图中可以直观地看到每个设计所能达到的计算峰值、CTC 率以及当前系统各分支并行度，每个点代表的是各分支网络潜在

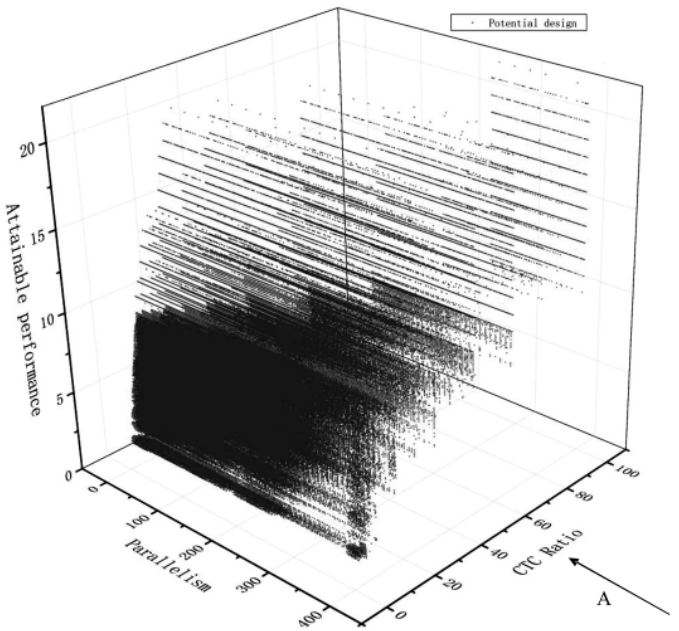

图 4　枚举模型

展开因子。所有的设计被 FPGA 平台所能提供的计算峰值和带宽所限制，即图中的计算屋顶和带宽边界。

由于本次设计的卷积计算单元是各卷积层之间通用的，即各分支的卷积层重复调用其特定的计算单元，因此需要统一的拆分因子。根据以上提出的方法，表 1 中列出了对于每个层来说最优的解。根据执行的周期数，在本设计中对所有的有效设计进行了枚举以找到优化的全局拆分因子。

3 系统架构

图 5 是所设计的基于 FPGA 的多支卷积神经网络加速器系统。由于 FPGA 芯片的片上 SRAM 不足以存储所有权重及输入输出数据，因此采用 DDR 和片上存储器的两级存储器层次结构，DDR4 DRAM 作为外部存储器，用来存储 CNN 的输入输出数据及权重。

表 1　每层最优解及统一拆分因子

	第一支拆分因子	第二支拆分因子	第三支拆分因子	执行周期
第一层	T_m=16，T_n=1	T_m=16，T_n=1	T_m=16，T_n=1	4 096
第二层	T_m=24，T_n=4	T_m=24，T_n=4	T_m=24，T_n=8	512
第三层	T_m=32，T_n=2	T_m=32，T_n=2	T_m=32，T_n=8	192
总计	–	–	–	4 800
统一拆分因子	T_m=32，T_n=2	T_m=32，T_n=2	T_m=32，T_n=8	4 800

通用 CPU 用于初始化基于 FPGA 的加速器和执行时间的测量，通过 AXI4Lite 总线配置 FPGA 逻辑中的 DMA，从而调度加速器的运行，控制数据传输的大小与格式。FPGA 中的 DMA 则通过 AXI4 总线读取外部存储 DDR 中的数据。为了提高数据的传输速率，DMA 与加速器之间的接口为 FIFO，进行数据的顺序访问，无需地址的判断。由于是三支网络同时进行计算，因此并行了三个计算单元，UART 模块将加速器返回的结果传送到主机。

4 实验结果

为了评估本文的优化策略，通过前面介绍的 ETH-CNN 模型在 FPGA 平台上构建多分支 CNN 加速器，并在 FPGA 上实现。系统设计使用 Xilinx 公司的 ZCU102 开发板进行验证，芯片型号为 XCZU9EG-2FFVB1156，FPGA 的工作频率为 200 MHz。作为对比，PC 平台使用的是 Intel i5-8300H CPU，主频为 2.3 GHz，并采用相同的网络结构及测试数据进行仿真验证。实验结果如表 2 所示。

表 2　计算时间对比

平台	时间/μs
Intel i5-8300H CPU	251
本文设计 XCZU9EG	25.22

最终的多分支 CNN 加速器包含 3 个加速单元，各单元并行计算不同分支的卷积输出。输入数据、输出数据以及权重数据的接口综合为使用基于 AXI 总线的 HP 高性能接口，采用的是 Scatter-gather 模式的 DMA。当三支并行加速时，总的资源利用率如表 3 所示。

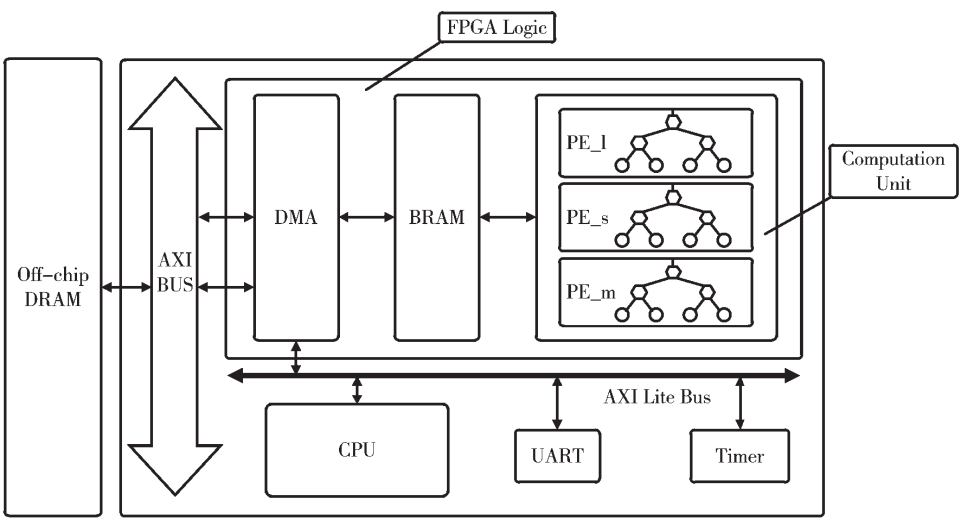

图 5　加速器整体架构

表 3 多分支卷积神经网络加速器 FPGA 资源利用率

资源	DSP	BRAM	FF	LUT
已用	2 217	404	332 618	247 624
可用	2 520	1 824	548 160	274 080
利用率/%	88	22	61	90

此外，由于 ETH-CNN 模型较小，运算量约为 0.56 MOP，因此在 FPGA 上实现的吞吐量有限。因此，也通过 GoogLeNet 的 Inception3a 结构来评估本文提出的多分支加速器映射方式，并与先前工作进行了比较。LCP[18]通过模型参数和数据位置差异将其按层簇的方式进行划分，并行于 FPGA 的不同分区实现加速。本文方法以支为界限划分，将不同分支的卷积神经网络映射至 FPGA 上并行执行，通过资源互补，采取不同的展开因子来配比各分支网络。从表 4 中可以看出，在 32 bit 的测试基准下，本文的工作取得了更有效的优化，取得了 257.01 GOPS 的性能表现，性能为先前方法的 1.31 倍。

表 4 性能评估

	文献[16]	本文	
计算精度	32 bit	32 bit	32 bit
FPGA 型号	VX690T	XCZU9EG	XCZU9EG
DSP 资源	2 916(81%)	2 217(88%)	2 203(87%)
FPGA 工作频率/MHz	200	200	200
计算操作量	128M	0.56M	128M
峰值性能/GOPS	196.27	22.17	257.01

4 结论

本文提出了一种基于 FPGA 的多分支 CNN 加速器的综合设计及优化方法。在给定多分支 CNN 模型和 FPGA 平台下，鉴于加宽 CNN 的结构导致的计算量增大，数据流更为复杂，针对并行网络进行更仔细的优化以及调度。利用屋顶模型进行了设计空间的探索，充分运用 FPGA 中的资源，进行并行性设计和流水线设计。实验结果表明，本文的工作相比先前方法进一步提升了 31%，取得了更有效的优化。

参考文献

[1] KHAN A, SOHAIL A, ZAHOORA U, et al. A survey of the recent architectures of deep convolutional neural networks[J]. Artificial Intelligence Review, 2019(1-87).

[2] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]. IEEE Conference on Computer Vision & Pattern Recognition, IEEE Computer Society, 2016.

[3] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]. IEEE Conference on Computer Vision & Pattern Recognition, IEEE Computer Society, 2015.

[4] KAWAGUCHI K, HUANG J, KAELBLING L P. Effect of depth and width on local minima in deep learning[J]. Neural Computation, 2019, 31(7): 1462-1498.

[5] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 1-9.

[6] DEL COCO M, CARCAGNI P, LEO M, et al. Multi-branch CNN for multi-scale age estimation[C]. International Conference on Image Analysis and Processing, 2017: 234-244.

[7] JIANG N, XU Y, ZHOU Z, et al. Multi-attribute driven vehicle re-identification with spatial-temporal re-ranking[C]. 2018 25th IEEE International Conference on Image Processing(ICIP). IEEE, 2018: 858-862.

[8] HU Y, LU M, LU X. Driving behaviour recognition from still images by using multi-stream fusion CNN[J]. Machine Vision and Applications, 2019, 30(5): 851-865.

[9] 徐喆，王玉辉. 基于候选区域和并行卷积神经网络的行人检测[J]. 计算机工程与应用，2019，55(22)：11-18，162.

[10] 卢丽强，郑思泽，肖倾城，等. 面向卷积神经网络的 FPGA 设计[J]. 中国科学(信息科学)，2019，49(3)：277-294.

[11] ZHANG C, LI P, SUN G, et al. Optimizing fpga-based accelerator design for deep convolutional neural networks[C]. ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, 2015.

[12] MA Y, CAO Y, VRUDHULA S, et al. Optimizing loop operation and dataflow in FPGA acceleration of deep convolutional neural networks[C]. Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays, 2017.

[13] MOTAMEDI M, GYSEL P, AKELLA V, et al. Design space exploration of FPGA-based deep convolutional neural networks[C]. 2016 21st Asia and South Pacific Design Automation Conference(ASP-DAC). IEEE, 2016.

[14] WILLIAMS S, WATERMAN A, PATTERSON D A. Roofline: an insightful visual performance model for multicore architectures[J]. Communications of the ACM, 2009, 52(4): 65-76.

[15] XU M, LI T, WANG Z, et al. Reducing complexity of HEVC: a deep learning approach[J]. IEEE Transactions on Image Processing, 2018, 27(10): 5044-5059.

[16] LIN X, YIN S, TU F, et al. LCP: a layer clusters paralleling mapping method for accelerating inception and residual networks on FPGA[C]. The 55th Annual Design Automation Conference, 2018.

(收稿日期：2021-01-06)

作者简介：

谢思璞(1996-)，男，硕士研究生，主要研究方向：硬件加速器设计与优化。

魏榕山(1980-)，男，博士，教授，主要研究方向：集成电路设计研究。

扫码下载电子文档

版权声明