光学 精密工程
OPTICS AND PRECISION ENGINEERING
核心期刊
主管：中国科学院
主办：中国科学院长春光学精密机械与物理研究所
中国仪器仪表学会
主编：曹健林

基于高层次融合的卷积神经网络FPGA硬件加速

魏楚亮，陈儒林，高谦，孙正隆

引用本文:
魏楚亮, 陈儒林, 高谦, 等. 基于高层次融合的卷积神经网络FPGA硬件加速[J]. 光学 精密工程, 2020, 28(5): 1212-1219.
WEI Chu-liang, CHEN Ru-lin, GAO Qian, et al. FPGA-based hardware acceleration for CNNs developed using high-Level synthesis[J]. *Optics and Precision Engineering*, 2020, 28(5): 1212-1219.

在线阅读 View online: https://doi.org/10.3788/OPE.20202805.1212

文章编号 1004-924X(2020)05-1212-08

基于高层次融合的卷积神经网络 FPGA 硬件加速

魏楚亮[1],陈儒林[1*],高 谦[2,3],孙正隆[2,3]
(1. 汕头大学 电子工程系,广东 汕头 515063;
2. 深圳市人工智能与机器人研究院,广东 深圳 518054;
3. 香港中文大学(深圳) 理工学院,广东 深圳 518172)

摘要:为了解决神经网络前向传播过程中的硬件加速问题,设计了一套基于 FPGA 编程工具 Vivado HLS 开发的 AlexNet 神经网络前向传播硬件加速系统。该系统能够确保在达到相关应用要求的基础上,有效地节省开发时间并降低开发成本。系统基于高级计算机语言 C++进行 FPGA 电路的仿真与开发,同时,灵活运用具有很高便捷性及可靠性的 Vivado HLS 中的 PIPELINE 和 ARRAY_PARTITION 指令进行系统优化。实验结果表明,AlexNet 神经网络在本文所构建的 FPGA 加速系统上的运行时间为 21.95 ms,比在传统 GPU 平台上的运行时 70 ms 少,运行速度要 3 倍以上。此外,每一层的网络都实现了分开封装操作,使系统可便捷地移植到其它成熟的卷积神经网络上,加速了深度学习在各类人工智能系统上的应用,在智能产业具有广泛的应用价值。

关 键 词:深度学习;现场可编程门阵列;高层次融合;硬件加速电路
中图分类号:TP18;TP391.4 **文献标识码**:A **doi**:10.3788/OPE.20202805.1212

FPGA-based hardware acceleration for CNNs developed using high-Level synthesis

WEI Chu-liang[1], CHEN Ru-lin[1*], GAO Qian[2,3], SUN Zheng-long[2,3]

(1. *Department of Electronic Engineering, Shantou University, Shantou* 515063, *China*;
2. *Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen* 518054, *China*;
3. *School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen* 518172, *China*)
* *Corresponding author, E-mail*: 16*rlchen@stu.edu.cn*

Abstract: To accelerate the forward-propagation process of deep-learning networks, a field-programmable gate array (FPGA) hardware-acceleration system for AlexNet was developed using Vivado High-Level Synthesis (HLS), which can greatly reduce the FPGA development cost. Using Vivado HLS, developers can design hardware architectures on an FPGA platform using C/C++ code instead of a hardware-description language. We implemented AlexNet on an FPGA platform using the

收稿日期:2019-12-10;**修订日期**:2020-02-03.
基金项目:广东省普通高校特色创新项目(No. 2018KTSCX061);揭阳市科技计划项目(No. 2019007&2019065);广东省重大科技专项(No. 2015B020233018);深圳市人工智能与机器人研究院项目(No. 2019-INT010)

HLS tool, and then used the PIPELINE and ARRAY_PARTITION directives to optimize the proposed system. An evaluation of the proposed system shows that its performance is three times better than a traditional computing-platform graphics processing unit (GPU). In the future, owing to the high-level encapsulation, the developed system can be easily transformed into other convolutional neural networks for practical operation, which shows its great portability and practical application value.

Key words: deep learning; Field Programmable Gate Array (FPGA); high level synthesis; hardware acceleration circuits

1 Introduction

In recent years, Convolutional Neural Networks (CNN) have become an important tool in certain informat-ics or engineering fields, e. g., computer vision[1-3], signal processing[4-5], and robotics[6-7], which require a complex artificial intelligence. Other complicated interdisciplinary applications[8-9], including stock-price prediction, gas exploration, medical imaging, etc., are also in need of CNNs.

Graphics Processing Units (GPUs) have been widely used as accelerators for CNNs. Potluri et al.[10] proposed a real-time discrete-time CNN system using a GPU developed with the Open Computing Language (Open CL); it showed better computing performance than the central processing unit (CPU). In addition, Strigl et al.[11] presented a CNN acceleration framework, based on a GPU, for complex problems, e. g., Optical Character Recognition (OCR) or face detection. Other works, including car-plate recognition[12] and denoising prior to image restoration[13], have been proposed using GPUs. GPUs have been proven to perform two to 24 times faster than CPUs.

The Field Programmable Gate Array (FPGA), a more powerful hardware-acceleration circuit, has a smaller clock-cycle requirement than a GPU for the same tasks[14] because of its richer embedded resources, e. g., Digital Signal-Processing (DSP) blocks, registers, and first-in-first-out queues (FIFOs)[15]. Zhang et al.[16] presented an FPGA-based accelerator for a CNN, which achieved a peak performance of 61.62 billion Floating-point Operations Per Second (GFLOPS) under a 100-MHz working frequency, and prominently outperformed the other implementations. However, the GPU is widely used as a deep-learning computing platform because of its efficient development process, while few developers choose FPGAs. According to Ref. [14], it took one person (postdoctoral level) two months to develop a GPU-based real-time phase-based optical-processing system, while it took two people (postdoctoral level) 15 months to finish the same system on an FPGA.

With the development of High-Level Synthesis (HLS), Xilinx presented a novel tool, Vivado HLS[17], to design large-scale complex FPGAs using high-level computer languages[18]. Traditionally, developers have needed to use inefficient, high-cost, low-level Hardware Description Languages (HDLs) for FPGA designs. Using Vivado HLS, developers use C/C++ instead of HDLs to design the FPGA architecture; then, the designed C/C++ code can be automatically converted to a Register-Transfer Level (RTL) model and HDL. Furthermore, Vivado HLS provides different directives to optimize the FPGA design to reduce the system latency and interval. It also shows the design evaluation.

In this paper, we developed an FPGA-based hardware-acceleration system for a CNN, which can be used in a real-time processing system. The rest of the paper is organized as follows. Section 2 introduces the AlexNet architecture.

Section 3 illustrates in detail how to develop AlexNet on an FPGA using the HLS tool and optimize the original model through optimization directives. A computing-performance comparison between the proposed FPGA system and a GPU platform is detailed in Section 4. Section 5 gives a forward-propagation test, based upon the proposed FPGA system. Finally, Section 6 presents a brief conclusion and a challenging project plan.

2 CNN architecture

Here, we chose AlexNet as the deep-learning model to test. AlexNet is widely used in computer-vision tasks[19-21] because of its reasonable trade-offs between speed and accuracy. The complete network comprises eight layers with training weights: the first five are convolution layers and the last three are fully connected. A Rectified Linear Unit (ReLU) non-linearity was implemented to follow every convolutional and fully-connected layer. Moreover, AlexNet has two normalization layers and three max-pooling layers. The author used a softmax function at the end of the network to distribute the different class labels. If we use ImageNet as a dataset to train the network, with every image having 227×227×3 pixels, the output will be a 1000-way one-dimensional vector because this dataset contains 1000 different classes. The overall AlexNet architecture and detailed information on each layer are shown in Tab. 1.

Tab. 1 AlexNet architecture

Layer name	Layer type	Details
Conv 1	96 11×11×3 Convolution	Stride [4 4] Padding [0 0 0 0]
Conv 2	256 5×5×48 Convolution	Stride [1 1] Padding [2 2 2 2]
Conv 3	384 3×3×256 Convolution	Stride [1 1] Padding [1 1 1 1]
Conv 4	384 3×3×192 Convolution	Stride [1 1] Padding [1 1 1 1]
Conv 5	256 3×3×192 Convolution	Stride [1 1] Padding [1 1 1 1]
FC 1	4096 fully-connected layer	
FC 2	4096 fully-connected layer	
FC 3	1000 fully-connected layer	

3 HLS-based development process

Traditionally, an FPGA can be developed at either the Gate Level (GL) or the Register-Transfer Level (RTL). Designing an FPGA in the traditional manner requires the developer to arrange a logic-gate circuit to satisfy the desired need. Many details must be considered, e. g., bit width and time sequence, which requires extensive development time, even for an experienced developer. According to Ref. [14], which compared the development cost of a GPU and a traditionally developed FPGA, the FPGA was much more complex than the GPU.

To reduce FPGA development costs and meet the requirements of more complicated computing tasks, the hardware should be designed at the algorithmic level, which means developers need only focus on the high-level specifications of the problem. For this reason, Xilinx produced Vivado, a new FPGA-development kit, for synthesizing and analyzing HDL architectures. One of Vivado's most important tools is HLS, which accepts synthesizable subsets of ANSI C/C++, SystemC, and Matlab. The code is analyzed and automati-cally converted into an RTL model and an HDL, which is traditionally generated by gate-level logic-synthesis development software.

Figure 1 shows the workflow for the FPGA development of AlexNet using Vivado HLS. In this system, we used C/C++ as the development language and set all of the computations to use a single floating-point data type. First, we designed AlexNet using a high-level language

(C/C＋＋) and conducted simulation experiments. Once the experimental results met our requirements, the C/C＋＋ code was converted to HDL and the RTL model was automatically generated through HLS. Furthermore, Vivado HLS provides C/RTL co-simulation to simulate different FPGA on-chip environments and evaluate the use of logic-gate resources in the proposed system.

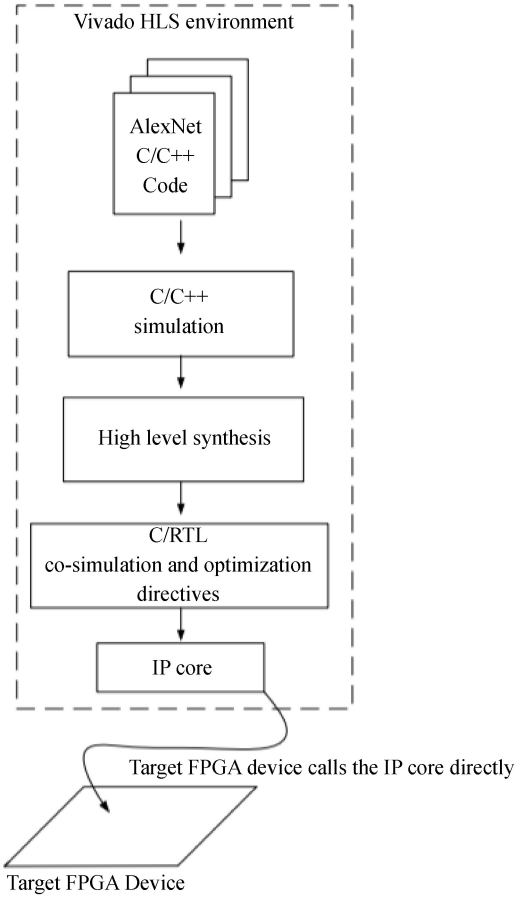

Fig. 1　Development workflow for AlexNet on an FPGA

To optimize the FPGA design, HLS has different directives that reduce the latency and interval. An optimization directive in HLS is another powerful tool to help developers design an FPGA at the algo-rithmic level. It can produce a micro-architecture that meets the desired requirement and area goals. We applied the PIPELINE and ARRAY _ PARTITION directives here. Through the PIPELINE directive, the next execution can start before the current execution has finished, which greatly reduces the initiation interval. The ARRAY _ PARTITION directive can partition large arrays into multiple smaller arrays or into individual registers, improving the access to data and removing block-RAM bottlenecks, which helps to reduce the latency. Figure 2 shows an example of using the optimization directives in Vivado HLS.

After optimization, the proposed system can be encapsulated into an intellectual property (IP) core. We can directly call the IP core from the FPGA development platform to complete the process of developing an FPGA through HLS, from the C/C＋＋ program to the FPGA on-chip system.

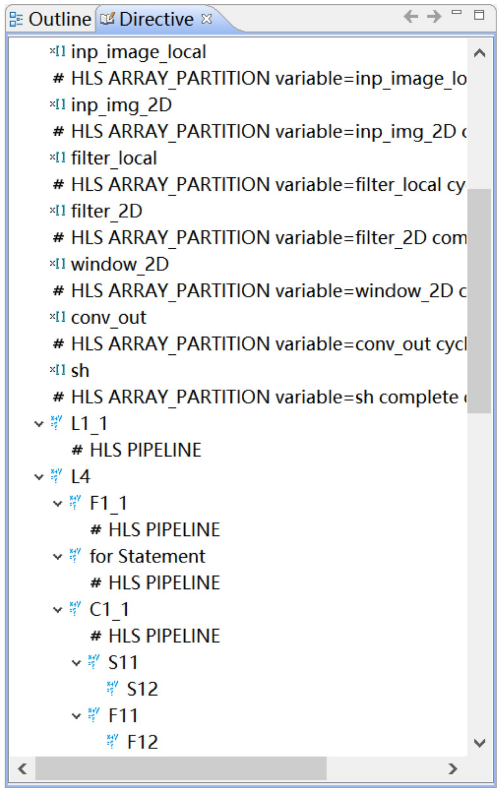

Fig. 2　Using optimization directives in one convolution layer

4　System-performance comparison

The proposed system implemented a pre-

trained AlexNet model with 60.5 k parameters on a Xilinx xcvu9p-flgb2104-2-i FPGA device. and the development environment was Vivado 2017.4. The operating frequency was set to 100 MHz. For comparison, we implemented the same model with the same parameter bit width in the an NVIDIA 960 m GPU with a 12-GB- memory working environment, developed by using Matlab 2018b.

The performance comparison between the FPGA and GPU platforms is shown in Figure 3. It took 21.95 ms for the proposed FPGA system to complete the forward-propagation procedure for a 227×227×3 pixel image. It took 70 ms on the traditional GPU platform. Thus, the computing speed on the FPGA platform is over three times faster than the GPU one.

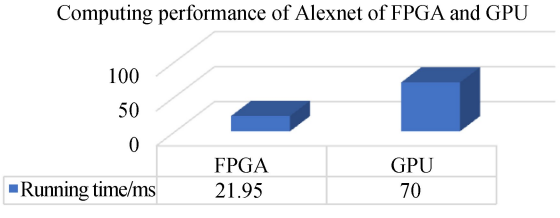

Fig. 3 Performance comparison between the FPGA and GPU platforms

Moreover, the detailed running time of each layer is shown in Figure 4. The execution time decreased from the first to last convolution layers, because the number of parameters was reduced after every convolution layer. Although there were only three fully-connected layers, they took 63.93% of the entire execution time to perform, as shown in Figure 5. Table 2 indicates the resource utilization of the proposed system, which is within the limit of the chosen FPGA board.

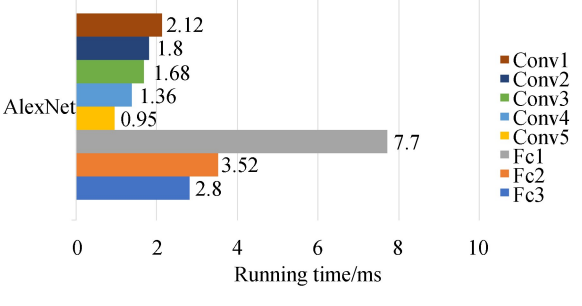

Fig. 4 Running time of each layer in AlexNet

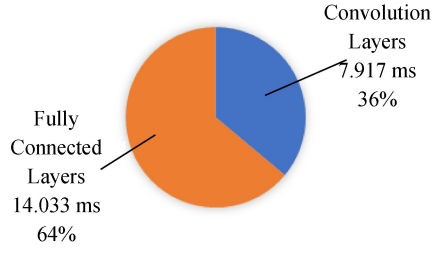

Fig. 5 Performance comparison between the convolution layers and fully-connected layers

Tab. 2 Resource utilization of Xilinx xcvu9p-flgb2104-2-i

Resource	Units utilized	Units available	Utilization
BRAM	1124	4320	26.01%
DSP	6686	6840	97.74%
FF	1404357	2364480	59.39%
LUT	1075078	1182240	90.93%

5 Forward-propagation test

To put the proposed FPGA system into practice, we used a tabby cat as one of our test inputs. It was obtained from the ImageNet database, which contains 1000 different classifications and was created by the Stanford Vision Lab, Stanford University. Figure 6 shows the input test image and the feature maps of each convolution layer, which indicates the successful forward-propagation process of the proposed FPGA system. With the forward propagation, the feature maps become less visually readable for human beings, but more mathematically understandable for the AlexNet model, as shown in Figures 6(b) to 6(f). Figure 7 shows the predic-

tion results of the input image after the three fully-connected layers and a softmax function. The successful implementation of the forward-propagation test proves that the system can be further used in other related tasks.

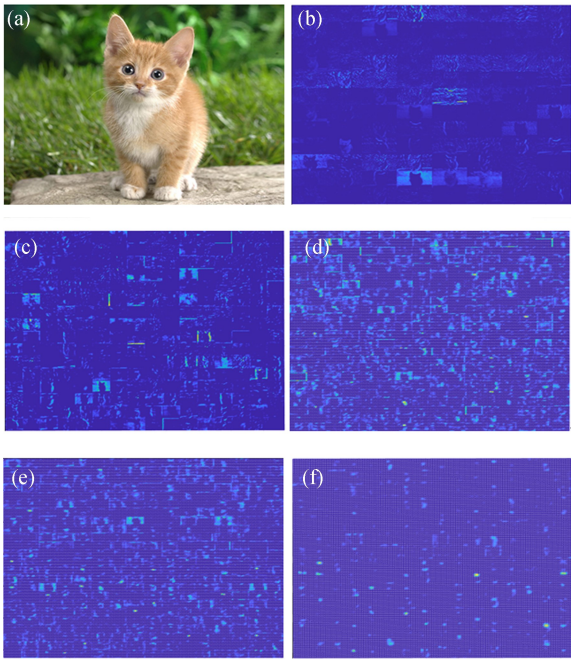

Fig. 6　(a) Input test image; (b) to (f) output feature maps of each convolution layer

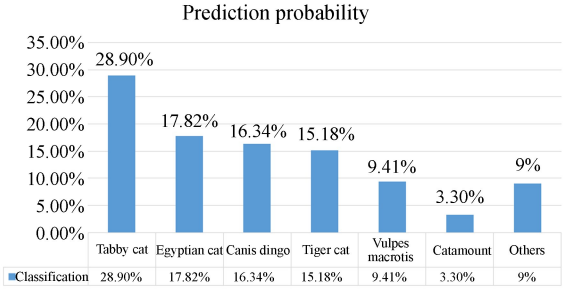

Fig. 7　Prediction probability results of the cat-image test

As future work, the proposed FPGA-based AlexNet network system will be used for further studies. For example, a human-robot interaction system, consisting of a UR5 robot arm, Kinect camera, force sensor, and infrared sensor, will be built in our laboratory. The system's image-processing speed should be as fast as possible to make it more stable and sensitive. Due to its limited resources and fixed circuit design, a GPU is not as applicable to this specific task as an FPGA.

6　Conclusion

This paper proposed an FPGA-based hardware-acceleration system for a deep learning network. The novel Vivado HLS was used as the development tool, instead of a traditional HDL. It enabled designs at the algorithmic level to reduce the development cost. AlexNet was selected as the deep-learning model to test in the proposed system. In the evaluation, the system showed better performance than a GPU. The proposed system can be further employed in various practical projects, e. g. , human-robot interaction systems, self-driving cars, and optical signal processing, to accelerate the processing procedure, while dealing with large-scale complex input data. The system can be divided into separate layers, which means it can be simply and flexibly transformed into other similar convolutional neural networks and used in different application scenarios.

Acknowledgments

This work was supported by the Characteristic Innovation Project of Universities in Guangdong Province under Grant No. 2018KTSCX061, the Projects of the Jieyang Science and Technology Plan under Grant No. 2019007 and Grant No. 2019065, the Key Project of Guangdong Province Science and Technology Plan under Grant No. 2015B020233018, and Project No. 2019-INT010 from the Shenzhen Institute of Artificial Intelligence and Robotics.

References:

[1] OZA P, PATEL V M. Deep CNN-based Multi-task Learning for Open-Set Recognition[EB/OL]. 2019.

[2] ZHU Y K, URTASUN R, SALAKHUTDINOV R, *et al*.. segDeepM: Exploiting segmentation and context in deep neural networks for object detection[C]. 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), 7-12 *June* 2015, *Boston*, *MA*, *USA*. *IEEE*, 2015: 4703-4711.

[3] GIRSHICK R, DONAHUE J, DARRELL T, *et al*.. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. 2014 *IEEE Conference on Computer Vision and Pattern Recognition*, 23-28 *June* 2014, *Columbus*, *OH*, *USA*. *IEEE*, 2014: 580-587.

[4] ZHANG D. A novel in-loop filtering mechanism of HEVC based on 3D sub-bands and CNN processing [J]. *Signal*, *Image and Video Processing*, 2019: 1-9.

[5] YE H, LI G Y, JUANG B H. Power of deep learning for channel estimation and signal detection in OFDM systems[J]. *IEEE Wireless Communications Letters*, 2018, 7(1): 114-117.

[6] LIANG F, ZHANG C. Hardware oriented vision system of logistics robotics[C]. 2018 12*th IEEE International Conference on Anti-Counterfeiting*, *Security*, *and Identification* (*ASID*), 9-11 *Nov.* 2018, *Xiamen*, *China*. *IEEE*, 2018: 6-9.

[7] GAO X, ZHANG T. Unsupervised learning to detect loops using deep neural networks for visual SLAM system[J]. *Autonomous Robots*, 2017, 41(1): 1-18.

[8] SELVIN S, VINAYAKUMAR R, GOPALAKRISHNAN E A, *et al*.. Stock price prediction using LSTM, RNN and CNN-sliding window model[C]. 2017 *International Conference on Advances in Computing*, *Communications and Informatics* (*ICACCI*), 13-16 *Sept.* 2017, *Udupi*, *India*. *IEEE*, 2017: 1643-1647.

[9] LI Q, CAI W D, WANG X G, *et al*.. Medical image classification with convolutional neural network [C]. 2014 13*th International Conference on Control Automation Robotics & Vision* (*ICARCV*), 10-12 *Dec.* 2014, *Singapore*, *Singapore*. *IEEE*, 2014: 844-848.

[10] POTLURI S, FASIH A, VUTUKURU L K, *et al*.. CNN based high performance computing for real time image processing on GPU[C]. *Proceedings of the Joint INDS′11 & ISTET′11*, 25-27 *July* 2011, *Klagenfurt*, *Austria*. *IEEE*, 2011: 1-7.

[11] STRIGL D, KOFLER K, PODLIPNIG S. Performance and scalability of GPU-based convolutional neural networks[C]. 2010 18*th Euromicro Conference on Parallel*, *Distributed and Network-Based Processing*, 17-19 *Feb.* 2010, *Pisa*, *Italy*. *IEEE*, 2010: 317-324.

[12] LEE S, SON K, KIM H, *et al*.. Car plate recognition based on CNN using embedded system with GPU [C]. 2017 10*th International Conference on Human System Interactions* (*HSI*), 17-19 *July* 2017, *Ulsan*, *South Korea*. *IEEE*, 2017: 239-241.

[13] ZHANG K, ZUO W M, GU S H, *et al*.. Learning deep CNN denoiser prior for image restoration[C]. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), 21-26 *July* 2017, *Honolulu*, *HI*, *USA*. *IEEE*, 2017: 2808-2817.

[14] PAUWELS K, TOMASI M, DIAZ ALONSO J, *et al*.. A comparison of FPGA and GPU for real-time phase-based optical flow, stereo, and local image features[J]. *IEEE Transactions on Computers*, 2012, 61(7): 999-1012.

[15] WANG X F, SOTIRIOS G Z. Hera: A reconfigurable and mixed-mode parallel computing engine on platform FPGAs[C]. 16*th International Conference on Parallel and Distributed Computing and Systems* (*PDCS*). 2004.

[16] ZHANG C. Optimizing FPGA-based accelerator design for deep convolutional neural networks [C]. *Proceedings of the* 2015 *ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*. *ACM*, 2015.

[17] FEIST T. Vivado design suite[J]. *White Paper*, 2012, 5: 30.

[18] WEI C L, CHEN R L, XIN Q. FPGA design of real-time MDFD system using high level synthesis [J]. *IEEE Access*, 2019, 7: 83664-83672.

[19] ALMISREB A A, JAMIL N, DIN N M. Utilizing AlexNet deep transfer learning for ear recognition [C]. 2018 *Fourth International Conference on Information Retrieval and Knowledge Management* (*CAMP*), 26-28 *March* 2018, *Kota Kinabalu*, *Malaysia*. *IEEE*, 2018: 1-5.

[20] LU S Y, LU Z H, ZHANG Y D. Pathological brain detection based on AlexNet and transfer

learning[J]. *Journal of Computational Science*, 2019, 30: 41-47.

[21] WAJAHAT N. Classification of breast cancer histology images using ALEXNET[C]. *International Conference Image Analysis and Recognition*. *Springer*, *Cham*, 2018.

作者简介：

魏楚亮(1979－)，男，广东揭阳人，博士，副教授，2006 年于英国利物浦大学获得博士学位，主要从事人工智能、机器人、传感技术、智能工业控制、交通运输安全检测、FPGA 设计等的研究。E-mail：clwei@stu.edu.cn

通讯作者：

陈儒林(1998－)，男，广东茂名人，主要从事机器学习、FPGA 设计等的研究。E-mail：16rlchen@stu.edu.cn

（本栏目编辑：秦　思）