第 59 卷 第 2 期/2022 年 1 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

一种用于激光焊接参数运算的可配置型BP神经网络计算加速器

范博宇[1*]，史再峰[1,3**]，王哲[1]，李少雄[1]，罗韬[2]

[1]天津大学微电子学院，天津 300072；

[2]天津大学智能与计算学部，天津 300072；

[3]天津市成像与感知微电子技术重点实验室，天津 300072

摘要 人工神经网络在各类激光技术中有着广泛应用，但是传统的流水展开架构加速器无法处理激光焊接参数提取、激光诱导击穿光谱分析等计算任务所需的多种反向传播(BP)神经网络。本课题组基于Xilinx PYNQ-Z2开发平台设计并实现了一种面向激光焊接技术的BP神经网络可配置型计算加速器架构。采用可配置架构设计和复用运算单元互连的方式，硬件电路可拟合成多种BP网络结构，加速器具有灵活的可配置性；同时，采用基于多级缓存结构的数据读取方法，解决了加速器运算阵列在读入数据时因多次访问片外存储器而导致的读取速度的瓶颈。基于实际激光焊接参数数据集的计算结果表明，所设计的加速器可以高效地加速具有多种神经元数量的BP神经网络。与嵌入式处理平台相比，加速器的典型网络运算性能平均有10.5倍的提升，神经元数目超过100的大型网络运算性能有56.4倍的提升，并且处理速度优于改进前于同一平台实现的普通加速器。

关键词 机器视觉；工业光学计量；BP神经网络；人工神经网络加速器；现场可编程门阵列

中图分类号 TN431.2 **文献标志码** A **doi**：10.3788/LOP202259.0214001

A Configurable BP Neural Network Accelerator for Laser Welding Parameter Calculation

Fan Boyu[1*], Shi Zaifeng[1,3**], Wang Zhe[1], Li Shaoxiong[1], Luo Tao[2]

[1]*School of Microelectronics, Tianjin University, Tianjin* 300072, *China*;

[2]*College of Intelligence and Computing, Tianjin University, Tianjin* 300072, *China*;

[3]*Tianjin Key Laboratory of Microelectronic Technology for Imaging and Sensing, Tianjin* 300072, *China*

Abstract Artificial neural networks are widely used in different types of laser technologies. However, traditional accelerators based on the pipeline deployment architecture cannot manage various back propagation (BP) neural networks needed for different laser calculation tasks, such as the extraction of laser welding parameters and laser-induced breakdown spectroscopy analysis. Based on the Xilinx PYNQ-Z2 development platform, a configurable accelerator architecture-based BP neural network for laser welding technologies is designed and implemented herein. By introducing the configurable accelerator architecture and the interconnection of multiplexing operation units, the hardware circuit can be fitted to various BP network structures and the accelerator shows flexible configurability. Furthermore, the data reading method based on a multilevel cache structure is adopted, which addresses the bottleneck of reading speed. Experimental results show that the proposed accelerator can efficiently accelerate the

收稿日期：2021-01-29；**修回日期**：2021-02-06；**录用日期**：2021-03-09

基金项目：国家自然科学基金(62071326)

通信作者：*fanboyu@tju.edu.cn；**shizaifeng@tju.edu.cn

BP neural network with various types of neurons. Compared with the embedded processor platform, the typical network operation performance of the proposed accelerator improves by 10.5 times on average and the large network operation performance with more than 100 neurons improves by 56.4 times on average. The proposed accelerator is superior to the general accelerator, which is realized on the same development platform.

Key words machine vision; industrial optical metrology; BP neural network; artificial neural network accelerator; field-programmable gate array

1 引 言

在激光焊接参数优化、激光切割过程控制、激光表面强化控制以及激光诱导击穿光谱分析等激光应用场景中，常面临多种控制参数的非线性映射问题。如：在激光表面强化控制中，需要研究材料热导率以及作用在材料表面的激光功率密度等参数对材料表面温度的影响，以控制钢铁材料的相变过程为奥氏体转变为马氏体[1]；在钢/铝激光焊接过程中，需要建立焊接速度、离焦量等参数与熔深之间的映射关系，以实现对Fe/Al脆性金属间化物的抑制[2]。在传统方法中，常使用热传导理论结合有限元分析、有限差分等方法进行计算。但这些方法常忽略了大功率激光分布不理想、材料表面吸收系数随温度非线性变化等因素，或对其进行了简化处理，难以对激光技术中的参数优化起到直接的指导作用。现阶段，很多激光技术大体上还处在总结实验规律阶段，需要通过大量实验来确定参数与性能之间的关系。

激光技术中的多输入、多输出、非线性等特性，使得传统的有限元分析、有限差分等多目标优化方法很难对上述需要极大运算量的工作进行最优处理[3]。人工神经网络(ANN)对复杂非凹非线性函数具有优异的拟合能力，在上述领域具有较为重要的应用价值。但传统的基于中央处理器(CPU)的软件实现方式能耗比很难高于1 GOPS/W(GIGA operations per second/W，每秒十亿次计算/瓦)，而且其运行速度也不理想[4]。这是由于软件实现的串行计算方法在调度和运算中具有周期消耗高的特点。使用硬件加速器对人工神经网络进行硬件实现，可以短时间、低功耗地进行大量的突发乘加运算，解决软件实现的速度、功耗不理想等问题。因此，考虑到人工神经网络大量计算的需求，定制硬件加速器对于提高其性能非常重要[5]。现场可编程逻辑门阵列(FPGA)以其并行性和低功耗等特点比较契合人工神经网络加速器的实现[6]。2016年，Ma等[7]提出了一种模块化人工神经网络加速器结构，该结构在Altera公司的Stratix-V系列FPGA平台上可达到114.5 GFLOPs(gigantic floating-point operation per second，每秒十亿次浮点数)的吞吐率。同年10月，Sharma等[8]提出了通用加速器——DnnWeaver，该加速器包含一个通用计算PU(processing unit)，进一步提升了加速器的可配置能力；但其使用软件对网络进行拆分的方法无法匹配多种网络结构。相比传统的实现方式而言，上述人工神经网络加速器的运行效率较高，可显著提升人工神经网络的前向推断速度并降低功耗，但存在灵活性与通用性不足的缺陷。

本课题组提出了面向激光焊接等技术的可配置型反向传播(BP)人工神经网络加速器架构，该结构采用可配置模块化设计方法，实现了对多种BP网络的支持。同时，该结构采用了基于多级缓存的数据读取方式与并行计算方法，提高了加速器运算单元的工作效率。本研究可为面向激光技术的人工神经网络加速器的开发提供参考。

2 BP神经网络

2.1 激光技术中的人工神经网络分析

人工神经网络具有优异的分类和信息提取能力，在激光技术等各研究领域均有一定的应用价值。近年来国内外在各类激光技术中选取的人工神经网络及取得的成果如表1所示。

从表1中可以看出，人工神经网络除了在辅助激光测距方面[12]取得的效果不显著外，在增材制造控制、激光诱导击穿光谱分析、激光焊接或切割参数优化等方面均取得了显著效果，尤其是在激光焊接等参数的优化上表现出了良好的应用潜力。分析表1可知，使用的网络结构主要有卷积神经网络、细胞神经网络和BP神经网络三种结构。卷积神经网络具有信息提取能力强和信息局部性利用程度高的特点，但其参数量往往特别巨大，一般的加速器架构难以提供足够的片上存储资源来满足其对存储空间的需求，因此难以进行硬件实现。细胞神经网络是一种局部互联、双值输出的网络结构，但

表1 近年来激光技术中的典型人工神经网络应用

Table 1 Typical application of artificial neural network in laser technology in recent years

Literature	Application direction	Network model used	Result
Literature [9]	Laser induced breakdown spectroscopy	Radial basis function neural network	The accuracy of some elements is improved
Literature [10]	Optimization of laser cutting parameters	Convolution neural network	About 92% accuracy
Literature [11]	Laser additive manufacturing control	Alex net	Effectively used in the process of image segmentation process
Literature [12]	Analysis of laser ranging data	Deep neural network	It is impossible for network to find deep information from satellite laser ranging data
Literature [13]	Spectral analysis	BP neural network	The modeling effect is improved
Literature [14]	Color laser marking	BP neural network	The feasibility is demonstrated
Literature [15]	Intensity calibration of laser scanner	BP neural network	The system response time is effectively shortened
Literature [16]	Laser induced breakdown spectroscopy	BP neural network	The results are satisfactory
Literature [17]	Target detection of optical genetic laser projection system	Convolution neural	Highly accurate detection effect is realized
Literature [2]	Optimization of laser welding parameters	BP neural network	The design goals of high precision, high quality, and high stability are realized
Literature [18]	Optimization of laser welding parameters	BP neural network	The relative error is small and the effect is good
Literature [19]	Laser welding control	Cellular neural network	The algorithm complexity are reduced and the control rate is high
Literature[20]	Optimization of laser cutting parameters	BP neural network	It has achieved obvious success

由于连接其神经元的权重的非线性特性，该网络的硬件化更适合使用忆阻器等新型处理元件实现，而使用该元件的新型加速器架构尚处于探索阶段，也不适合现有的硬件实现。BP神经网络是目前应用得较广泛的网络结构，该网络由输入层、隐藏层和输出层组成，其主要的计算消耗在输入参数矩阵与输入层到隐藏层的参数矩阵的相乘，以及隐藏层参数矩阵与隐藏层到输出层参数矩阵的相乘上。此类乘加运算比较适合通过调用硬件单元的方法进行计算，所以比较适合硬件实现。因此，设计一种可配置BP神经网络的加速器，可以在一定程度上提升网络的实用价值。

2.2 BP神经网络结构的并行加速技术分析

BP神经网络主要由输入层、隐藏层、输出层组成。若给定广义点对训练集 $D=\{(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)\}$，记隐藏层到输出层的权重为 $w_{1j}, w_{2j}, \cdots, w_{qj}$，隐藏层到输出层的阈值为 θ_i，第 i 个神经元的输出为 y_i^k，则有

$$y_i^k = f\left(\sum_{i=1}^{n} w_{ij} b_i - \theta_i\right), \tag{1}$$

式中：$f(*)$ 指激活函数；b_i 为该处神经元收到的刺激强度。此结构可使神经元模型获得处理非线性问题的能力。在总体的 i 次乘加运算中，各次运算并没有逻辑上的依赖关系。为利用硬件电路的并行化特性，提高加速器的运算效率，本课题组设计了加速器的并行运算加速技术，具体过程如图1所示。

图1中，$v_{1h}, v_{2h}, \cdots, v_{nh}$ 是输入层到隐藏层的权重。加速器通过并行化运算单元阵列同时对多次乘加运算进行计算。通过在加速器的可编程逻辑端设计 k 个运算核心，可以实现并行度为 j 的加速计算（$j \leqslant k$）。多次调用该组运算核心可以计算出任

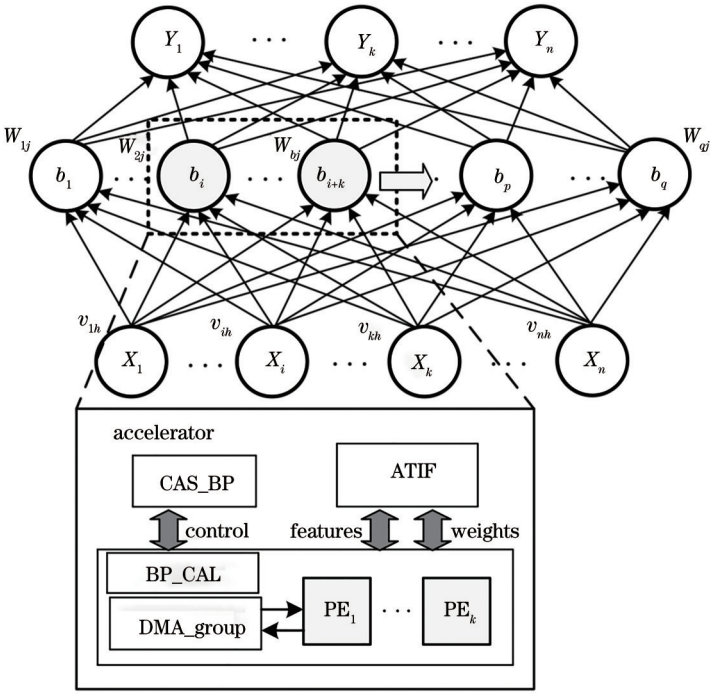

图 1　BP 神经网络并行加速技术

Fig. 1　Parallel acceleration technology of BP neural network

意神经元数目的 BP 神经网络的运算结果，从而达到拟合多种不同神经元数目的网络结构的目的。在实际设计的架构中，输入数据并行分别流入由 5 个处理元件（PE）组成的并行运算阵列中协同运算。该并行加速技术通过多处理元件协同计算提高了加速器的工作效率。理论上，并行度为 j 的并行计算方式最多可以提升 j 倍的运行效率，但由于引入了额外的控制逻辑以及数据读入速度等限制，实际设计中得到的效率提升要小于理论值。

3　可配置 BP 网络的加速器设计

3.1　模块化可配置架构

为使面向激光技术的 BP 神经网络加速器能够配置为多种应用场景下的不同 BP 神经网络，本课题组基于 Top-Down 设计思想提出了可配置模块复用的加速器架构，具体如图 2 所示。该架构主要由 DDR SDRAM（double data rate synchronous dynamic random access memory）控制模块、接口模块（ATIF）、BP 运算模块（BP_CAL）和参数配置模块（CAS_BP）组成。该加速器工作时，首先由嵌入式系统（PS）端的 ARM Cortex-A9 CPU 通过 AXI_lite 总线对加速器可编程逻辑（PL）端的配置模块进行发包，由接收状态机（receive FSM）配置内部控制寄存器组（command group）。配置完成后发送状态机（send FSM）向 PL 端的运算模块发送指令包。BP 运算模块由其内部的 DMA 组（direct memory access group）根据配置发出数据请求，并通过接口模块将请求转为 AXI_full 格式的通信协议，通过 DDR 控制模块从 DDR SDRAM 中取对应的权重和输入参数，送入并行 PE 组中进行运算，运算完成后反馈完成标志位，进行下一次配置与计算。通过这种多次配置多个运算核心的工作方式，可以根据 PE 组运算轮数将 BP 神经网络加速器拟合成任意神经元数量的 BP 神经网络。这种架构使加速器具有了灵活的可配置能力。

由于 DDR SDRAM 的周期性预充电特征，该存储单元无法支持连续读写，这会导致加速器工作时数据的读取速度显著慢于运算速度，进而导致整个系统的工作效率下降。针对这个问题，此 BP 神经网络加速器的架构中使用了分级数据存储模式，即：将数据按运算需求分批存入片内的多级存储结构中。由图 2 可以看出：DDR SDRAM 为 BP 神经网络权重数据存储单元；BP_CAL 模块左侧使用 BRAM（block RAM）作为片上缓存，存储当前运算所需的全部数据，而右侧则是由多个寄存器组成的阵列 PE 寄存器文件（RF）。RF 通常由静态随机读写存储器（SRAM）实现，可以多路并发并访问不同的地址，具有响应快的优点。现有研究表明，这种

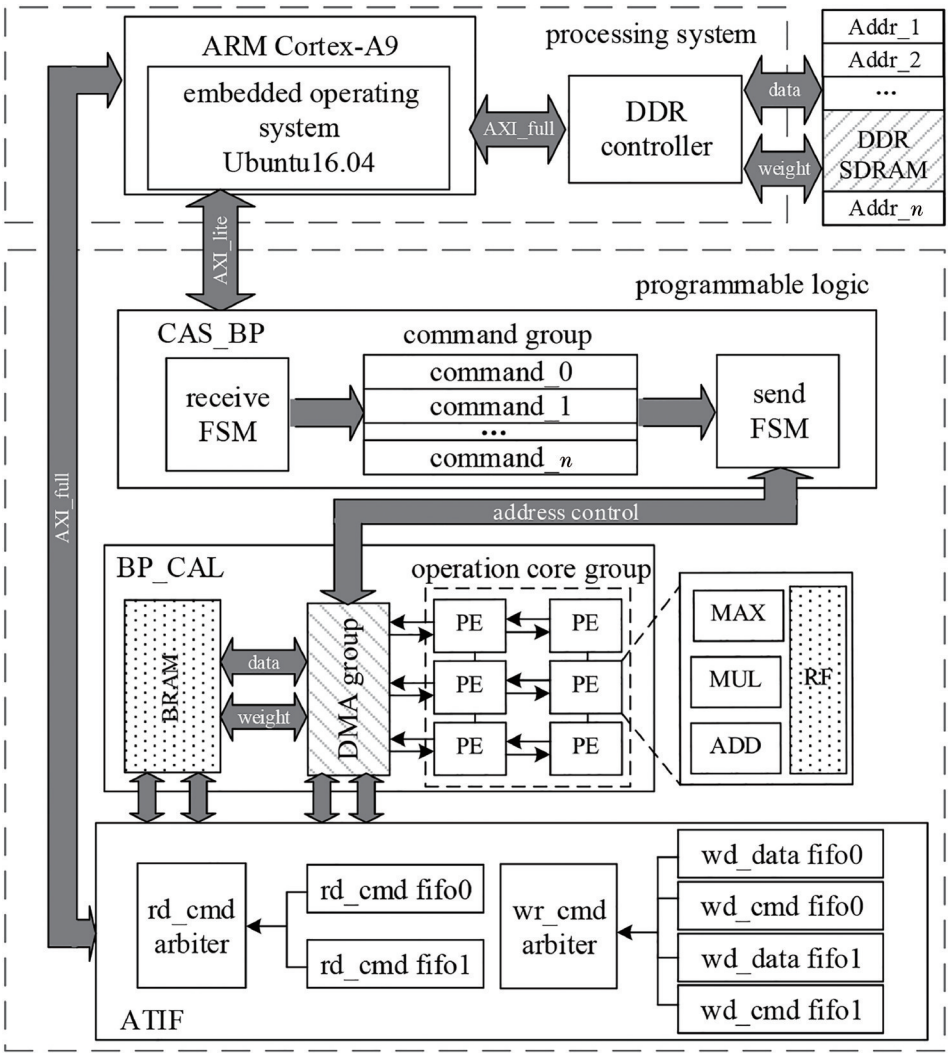

图 2　加速器整体架构与分级缓存结构

Fig. 2　Accelerator architecture and multi-level storage of data

分级结构可以有效提升数据的读取效率[21]。

3.2　多级缓存结构的数据读取方法

在处理包含较大数目隐藏层神经元的 BP 神经网络时，并行运算阵列通常在每笔局部运算结束后等待重新写入新的特征与权重矩阵，因此，整个系统面临着工作不连续的问题。本加速器中设计了基于乒乓 buffer 结构（分别为图 3 中的 Fbuffer_A、Fbuffer_B 与 Wbuffer_A、Wbuffer_B）的片上缓存以及相应的 DMA，以实现系统工作的连续性。具体过程如图 3 所示。

权重与特征数据存储在片外 DDR SDRAM 中，在运算至相应神经元时，由 FDMA（feature DMA）和 WDMA（weight DMA）这两个外部 DMA 模块负责将其从外部 DDR SDRAM 输入两组片上的 BRAM 缓存中。为了保证数据可以进行连续处理，每组片上的 BRAM 缓存由两块 BRAM 组成，使用乒乓的方式从其中读写数据。这种方式可以在很大程度上加快整个加速器的处理效率，减少运算矩阵的等待时间。上述 AB buffer 中数据的写入与传输在不同周期中轮流进行，数据传输至运算阵列时，由特征运算 DMA（feature operation DMA，FoDMA）和权重运算 DMA（weight operation DMA，WoDMA）两个内部 DMA 控制。两个 DMA 之间的控制逻辑通过共享权重和特征计数信号实现，保证了特征和权重的协调工作，避免了运算矩阵读入不匹配数据以及权重导致运算结果错误。由于运算速度有时可能会大于写入速度，buffer 的写位置控制信号同样需要接给内部 DMA，以保证读地址时不会覆盖到当次运算中还未写入新数据的地址。此多级缓存数据读取方法可以在保证运算正确性的同时实现数据的连续读入与运算。

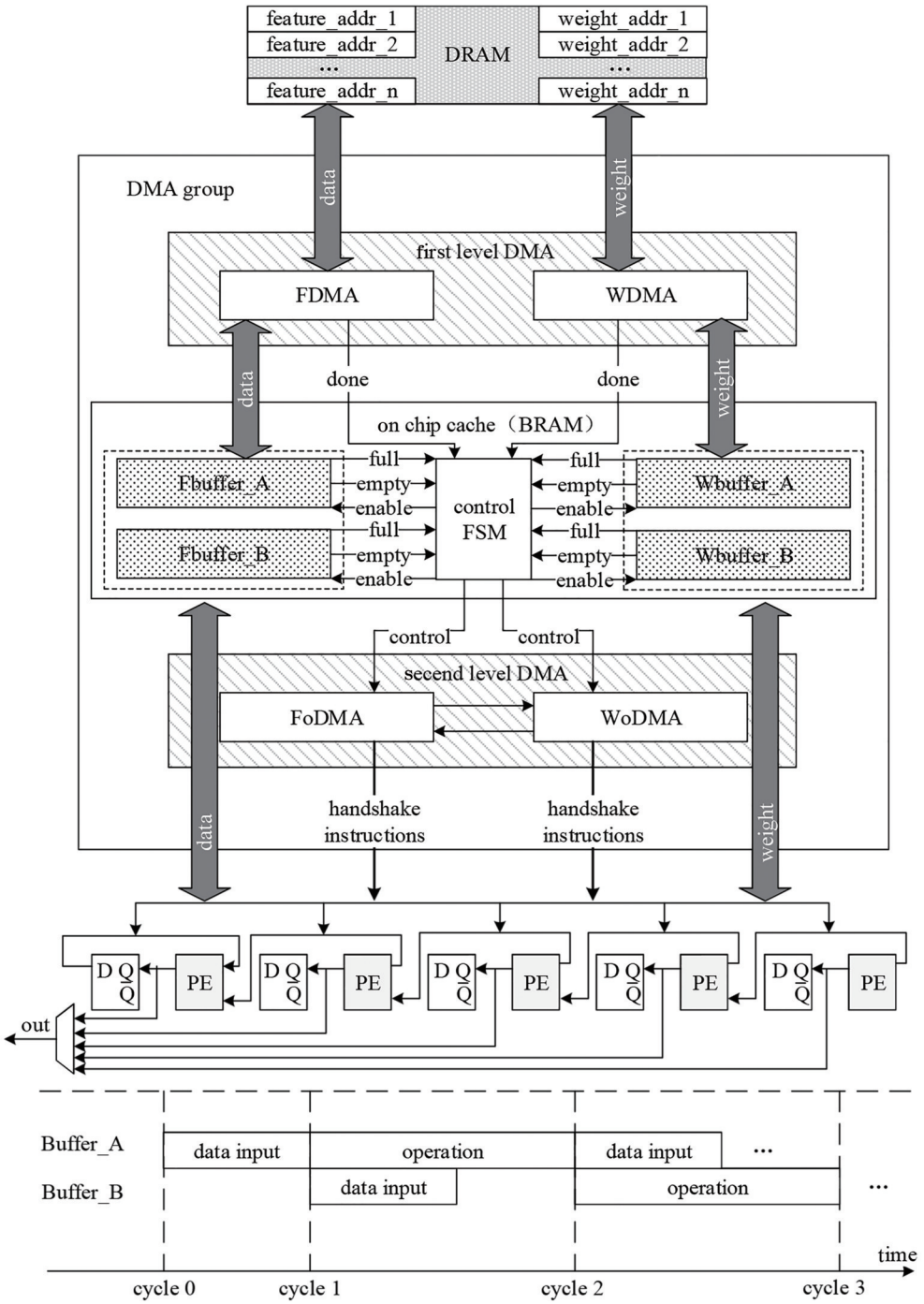

图 3　运算矩阵的数据传输模块

Fig. 3　Data transmission module of operation matrix

4　实验结果与分析

4.1　实验环境与验证平台

本设计使用 Xilinx 公司的 Zynq-7000 系列可编程片上系统(SOPC)作为实现平台，所使用的数据来自 Wedding0327 焊接参数数据集，其中包括了扫描速度、功率、离焦量等输入参数与焊缝宽度、熔深等输出参数的映射关系，典型参数映射关系如表 2 所示。

本文基于验证方法学(UVM)技术，使用 Python3 软件语言在开发板上搭建了验证平台，其结构如图 4 所示。其中，映射序列(sequence)在进行实际数据集运算时，可以直接输入待映射参数作为激励。上述激励组成输入矩阵，作为多个输入序列(transaction_i)，将其存入作为映射机(sequencer)的数组中。映射驱动(driver)函数将数据从映射机中

表 2　数据集中典型的参数映射关系

Table 2　Typical mapping relationship of parameters in data set

Scanning speed /(m·min^{-1})	Power /W	Defocus /mm	Weld width /μm	Penetration /μm	Ratio of penetration to weld width
1.3	1000	10	1037.3130	1985.92	1.914484
1.5	800	0	733.7967	1859.60	2.534217
1.5	1000	3	1482.8600	1989.30	1.341529
1.8	1000	0	819.4533	1976.73	2.412255
2.5	1500	3	1166.8770	1958.60	1.678498
3.0	1500	0	776.9867	1858.92	2.392473
5.3	2000	3	617.0433	1900.31	3.079703
5.5	2000	0	597.3000	1855.23	3.106027
7.0	2500	0	676.2867	1856.02	2.744428

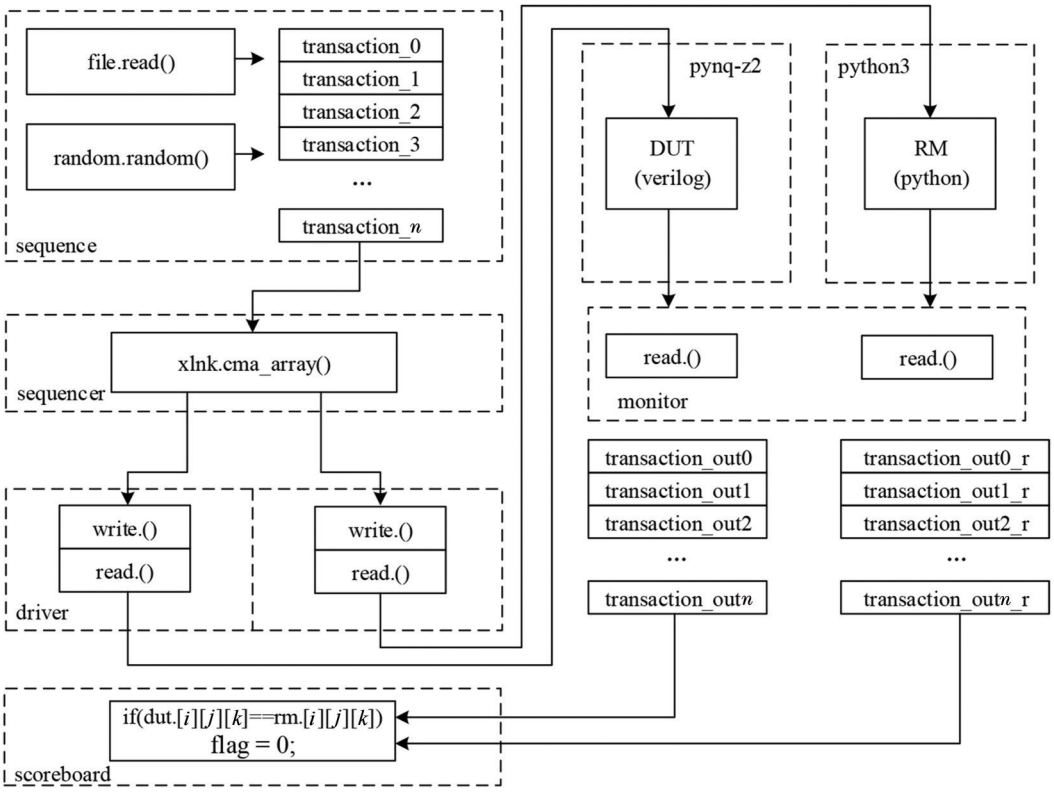

图 4　加速器专用验证平台结构

Fig. 4　Structure of our accelerator specific verification platform

取出，并分别输送给硬件待测设计(DUT)和行为级参考模型(RM)。二者的运算结果由显示模块(monitor)负责从结果寄存器中收集，然后分别存入缓存矩阵中并包装成两个输出序列(transaction_out$i$)，最终输送给计分板(scoreboard)进行比对并输出比对结果标志。在这个过程中，使用的参考模型是用Python3语言编写的加速器行为级模型。将所得结果与DUT的运算结果进行比较，以便对设计的BP神经网络硬件加速器进行验证。

4.2　实验结果与性能分析

加速器选用Xilinx公司的xc7z020clg400-1芯片进行设计实现。布局布线结果显示：硬件实现共消耗了9382个LUT(Look Up Table)，共占片上总资源的17.6%，同时消耗了12686个FF(Flip Flop)，占片上总资源的11.9%；23个数字信号处理器(DSP)占片上总资源的10.5%。整个系统片上功耗为1.578 W，总体功耗处在较低水平。

实验中使用的参考模型是使用Python3语言编

写的BP神经网络前向推断模型，它可以输出从输入层到隐藏层、隐藏层到输出层前向推断运算的所有结果。根据验证平台的设计原理，只有与参考模型输出比对一致的加速器测试结果才能得到统计数据。整个测试平台在Linux系统中运行，使用Python3语言进行硬件电路的驱动与配置。在实验中，本课题组对输入节点数量为3和4、隐藏层节点数量为4～10、输出节点数量为2～4等42种不同的网络结构分别进行了测试。加速器与嵌入式处理平台在典型情况下的处理速度对比如表3所示。从表3中表征神经元数目的Number of neurons列数据可以看出，加速器输入层、隐藏层、输出层中的神经元节点数目均可以任意配置。同时，从bp1处理时间（输入层到隐藏层的处理时间，t_{bp1}）、bp2处理时间（隐藏层到输出层的处理时间，t_{bp2}）列中的加速器处理速度与嵌入式处理平台处理速度的对比可以看出，所设计的加速器的处理速度显著快于嵌入式处理平台的处理速度。

表3　加速器处理速度对比

Table 3　Comparison of accelerator processing speed

Number of neurons			Processing time of bp1 t_{bp1} /ms		Processing time of bp2 t_{bp2} /ms	
Input layer	Hidden layer	Output layer	Accelerator	Embedded processor platform	Accelerator	Embedded processor platform
3	4	3	0.651360	3.217697	0.557184	2.809763
3	6	4	0.645399	4.746675	0.546455	5.123854
3	8	3	0.649691	6.266117	0.546694	4.949093
3	9	3	0.665188	7.077932	0.545502	5.591869
3	9	4	0.668287	6.937027	0.582457	7.375479
3	10	2	0.664234	7.764816	0.548601	4.140139
3	10	3	0.663519	7.673979	0.619888	6.249666
3	10	4	0.653505	7.800817	0.561714	8.289099
4	4	3	0.648975	4.116058	0.545502	2.835274
4	6	4	0.647545	5.889177	0.549316	5.116224
4	8	3	0.658512	7.790327	0.544310	5.586386
4	9	3	0.666857	8.561611	0.621557	5.743742
4	9	4	0.665903	8.777618	0.571012	7.631302
4	10	2	0.678539	9.542465	0.555515	4.715443
4	10	3	0.666857	9.572506	0.617504	6.770372

通过计算可知，与嵌入式处理平台相比，所设计的加速器相在运算典型网络时平均有10.5倍的速度提升，而且随着神经元数目的增加，速度提升的倍数呈现显著上升趋势。这说明加速器采取的并行运算策略有效提升了各层间神经元的计算速度，所设计的多级缓存数据读取方法可使运算阵列连续工作，获得较好的运算加速效果。

此外，为了使典型神经元数目下加速器的运算速度优势更加直观，将加速器运行时间与嵌入式处理平台的对比结果绘制在一幅图中，如图5所示。其中，上层时间曲面为嵌入式处理平台的运行时间统计，下层为加速器的运行时间统计，左侧条形图例对应嵌入式平台运行时间，右侧条形图例对应加速器运行时间。为了使图像更加清晰，对数据点进行了平滑处理。

从图5中可以看出，加速器的运行时间曲面整体上处于嵌入式处理平台运行时间曲面的下方。这表明，BP神经网络加速器的运算速度显著快于嵌入式处理平台的运算速度，证明了加速器加速的有效性。同时，来自多种神经元数目的性能数据证明了加速器的可配置性。

4.3　其他情况下的加速器性能分析

随着激光工业与BP神经网络技术的快速发展，其可能面对的输入参数也将不断增多。为了测试加速器在多输入、多隐含、多输出的复杂计算环境下的性能，本课题组还对包含16～128个神经元数量的BP神经网络分别进行了测试，测试结果如图6所示。图6(a)为输入神经元、隐藏层神经元、输出神经元在典型数量下的运行时间，图6(b)为较大神经元数目下的运行时间，虚线为加速器运行时间上界与嵌入式处理器平台运行时间下界的拟合曲

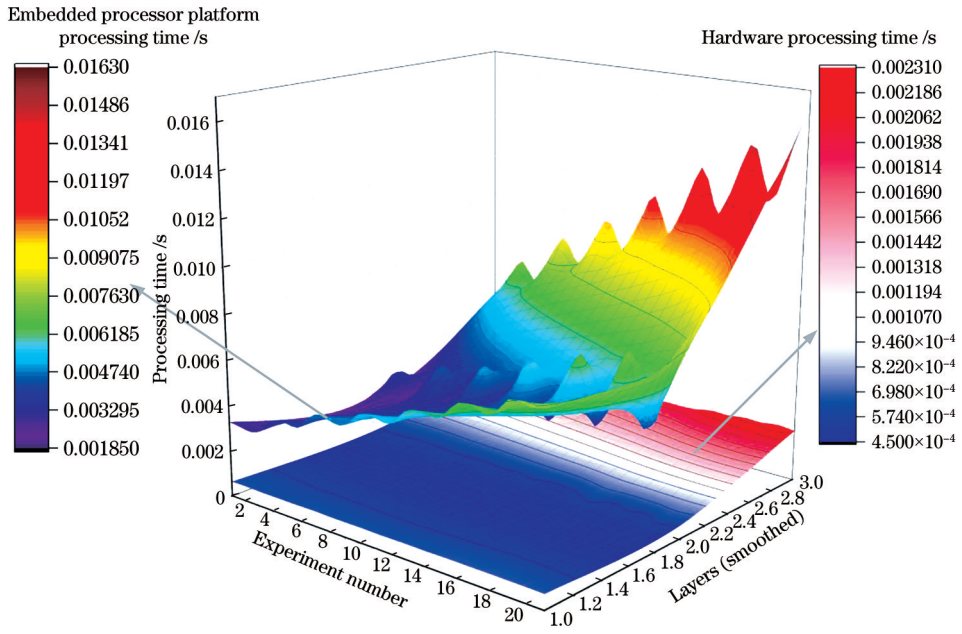

图 5　加速器与嵌入式处理平台的运行时间曲面对比

Fig. 5　Running time comparison between accelerator and embedded processor platform

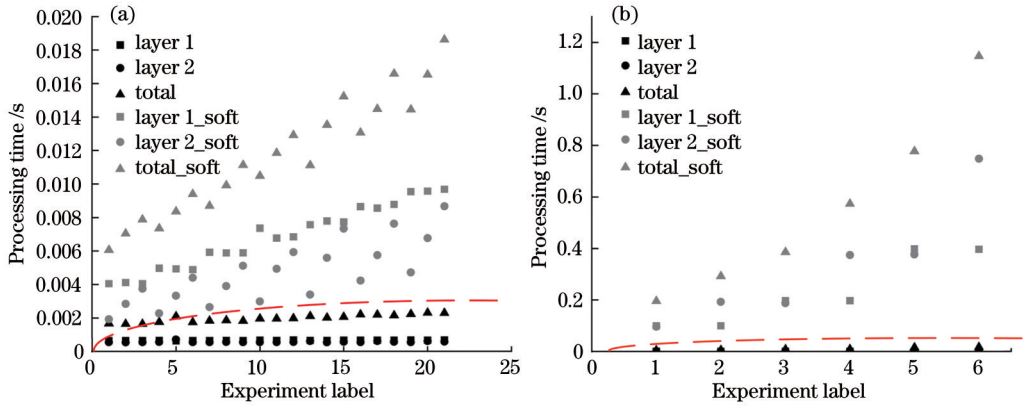

图 6　典型神经元数目下与较大神经元数目下加速器与嵌入式处理平台的运行时间对比图。(a)典型神经元数目；(b)较大神经元数目

Fig. 6　Running time comparison of accelerator and embedded processor platform under typical number neurons and maximum number of neurons. (a) Typical number of neurons; (b) maximum number of neurons

线。该拟合曲线与x轴越近，表示加速器的运行效果越好。从图 6(b)中可以看出，在较大神经元数目的运算中，加速器的运算表现相较于嵌入式处理平台在速度上的优越性更加显著，总处理速度约有 56.4 倍的提升。可以看出，在处理任务量较大的运算任务时，BP 神经网络加速器的运算效率优势更加明显。这反映了本文所提出的加速器架构在采取并行运算策略时的运算速度优势及可配置能力。

为体现本加速器设计方法的优势，本课题组还搭建了一个使用串行运算方法对数据进行计算的普通神经网络加速器，用于性能对比。为避免不同开发环境与设备性能的影响，该普通加速器使用相同的 FPGA 开发平台进行了布局布线与板上实现。实验结果如表 4 所示。

表 4 中，Proportion 列为本文所设计的加速器与普通加速器输入层到隐藏层处理时间(t_{bp1_a})与隐藏层到输出层处理时间(t_{bp2_a})的比值。从该列数据可以看出，本文所设计的加速器的处理速度显著快于改进前未使用并行计算方式设计的普通加速器，并且，随着神经元数目的增加，本文所设计的加速器的性能优势有显著的上升趋势。这一对比结果证明了本加速器架构在运算速度上的优势。

表 4 加速器间处理速度的对比

Table 4 Comparison of processing speed between accelerators

Number of neurons			Processing time of bp1 t_{bp1_a} /ms		Processing time of bp2 t_{bp2_a} /ms		Proportion
Input layer	Hidden layer	Output layer	General accelerator	Proposed accelerator	General accelerator	Proposed accelerator	
16	32	16	1.061678	0.909328	0.926495	0.851393	0.8877
16	32	32	0.994444	0.916243	1.302719	1.132965	0.8955
32	64	32	2.145052	1.786947	2.032518	1.715899	0.8386
32	64	64	2.478838	1.778364	3.809452	2.868414	0.7352
64	128	64	6.696224	5.238771	6.526232	5.165577	0.7869
64	128	128	6.678343	5.263567	12.558460	9.703398	0.7804

5 结 论

本课题组在对激光焊接等技术中的人工神经网络应用进行统计与分析的基础上，提出并设计了一种面向激光焊接参数计算的BP神经网络可配置型加速器架构，并对加速器进行了硬件实现。为使加速器可以加速多种不同神经元数目的BP神经网络，设计了加速器的模块化可配置复用架构；同时采用片外存储器、片上缓存、寄存器文件等存储器件开发了多级缓存数据读取方法，实现了运算的连续进行。基于激光焊接数据集的实验结果表明：所设计的可配置型加速器架构可以正确地将硬件电路配置为具有多种不同神经元数目的BP神经网络结构，实现了加速器配置的灵活性；开发的多级缓存数据读取方法实现了加速器并行运算阵列的连贯计算。与嵌入式处理平台相比，所设计的加速器的典型网络运算性能有10.5倍的提升，神经元数目超过100的大型网络运算性能有56.4倍的提升；与改进前的普通加速器架构相比，所设计的加速器运算速度优势明显。所设计的加速器具备一定的通用性，为面向激光技术的人工神经网络加速器设计提供了可行性思路。

参 考 文 献

[1] Wang D C. Controlling laser surface strengthening process based on artificial neural network[J]. Laser Technology, 2003, 27(4): 317-320.
王大承. 人工神经网络在激光表面强化控制上的应用[J]. 激光技术, 2003, 27(4): 317-320.

[2] Guo L, Wang S H, Zhang Q M, et al. Optimization of fiber laser welding process viarables and performance prediction based on BP neural network[J]. Applied Laser, 2010, 30(6): 479-482.
郭亮, 王少华, 张庆茂, 等. 基于BP神经网络的光纤激光焊接工艺参数优化及性能预测[J]. 应用激光, 2010, 30(6): 479-482.

[3] Bandoh S, Nakayama Y, Asagumo R, et al. Establishment of database of carbon/epoxy material properties and design values on durability and environmental resistance[J]. Advanced Composite Materials, 2002, 11(4): 365-374.

[4] Venieris S I, Bouganis C S. fpgaConvNet: a framework for mapping convolutional neural networks on FPGAs[C]//2016 IEEE 24th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM), May 1-3, 2016, Washington, DC, USA. New York: IEEE Press, 2016: 40-47.

[5] Mittal S. A survey of FPGA-based accelerators for convolutional neural networks[J]. Neural Computing and Applications, 2020, 32(4): 1109-1139.

[6] Zhao Y Q, Zhang X, Fang X, et al. A deep residual networks accelerator on FPGA[C]//2019 Eleventh International Conference on Advanced Computational Intelligence (ICACI), June 7-9, 2019, Guilin, China. New York: IEEE Press, 2019: 13-17.

[7] Ma Y F, Suda N, Cao Y, et al. Scalable and modularized RTL compilation of convolutional neural networks onto FPGA[C]//2016 26th International Conference on Field Programmable Logic and Applications (FPL), August 29-September 2, 2016, Lausanne, Switzerland. New York: IEEE Press, 2016: 1-8.

[8] Sharma H, Park J, Mahajan D, et al. From high-level deep neural models to FPGAs[C]//2016 49th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO), October 15-19, 2016, Taipei, Taiwan, China. New York: IEEE Press, 2016: 1-12.

[9] Pan L J, Chen W F, Cui R F, et al. Quantitative analysis of aluminum alloy based on laser-induced breakdown spectroscopy and radial basis function

neural network[J]. Laser & Optoelectronics Progress, 2020, 57(19): 193002.
潘立剑, 陈蔚芳, 文良华, 等. 基于激光诱导击穿光谱与径向基函数神经网络的铝合金定量分析[J]. 激光与光电子学进展, 2020, 57(19): 193002.

[10] Franceschetti L, Pacher M, Tanelli M, et al. Dross attachment estimation in the laser-cutting process via Convolutional Neural Networks (CNN) [C]//2020 28th Mediterranean Conference on Control and Automation (MED), September 15-18, 2020, Saint-Raphaël, France. New York: IEEE Press, 2020: 850-855.

[11] Elwarfalli H, Papazoglou D, Erdahl D, et al. *In situ* process monitoring for laser-powder bed fusion using convolutional neural networks and infrared tomography [C]//2019 IEEE National Aerospace and Electronics Conference (NAECON), July 15-19, 2019, Dayton, OH, USA. New York: IEEE Press, 2019: 323-327.

[12] Xue L, Zhu Z K, Wu W T, et al. Simulated analysis of processing satellite laser ranging data using neural networks trained by DeepLabCut[C]//2019 IEEE 5th International Conference on Computer and Communications (ICCC), December 6-9, 2019, Chengdu, China. New York: IEEE Press, 2019: 468-472.

[13] Hu J, Liu Y D, Sun X D, et al. Quantitative determination of benzoic acid in flour based on terahertz time-domain spectroscopy and BPNN model [J]. Laser & Optoelectronics Progress, 2020, 57(7): 073002.
胡军, 刘燕德, 孙旭东, 等. 基于BP神经网络的太赫兹时域光谱对面粉中苯甲酸的定量检测研究[J]. 激光与光电子学进展, 2020, 57(7): 073002.

[14] Cacivkins P, Lazov L, Teirumnieks E, et al. Artificial neural networks: what can they learn about color laser marking? [C]//2018 IX National Conference with International Participation (ELECTRONICA), May 17-18, 2018, Sofia, Bulgaria. New York: IEEE Press, 2018: 1-4.

[15] de Figueiredo R M, Veronez M R, Tognoli F M W, et al. Laser scanner intensity calibration based on artificial neural networks[C]//2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), July 23-28, 2017, Fort Worth, TX, USA. New York: IEEE Press, 2017: 1716-1719.

[16] Hu Y, Li Z H, Lü T. Quantitative measurement of iron content in geological standard samples by laser-induced breakdown spectroscopy combined with artificial neural network[J]. Laser & Optoelectronics Progress, 2017, 54(5): 053003.
胡杨, 李子涵, 吕涛. 激光诱导击穿光谱结合人工神经网络测定地质标样中的铁含量[J]. 激光与光电子学进展, 2017, 54(5): 053003.

[17] Shi Z F, Ye P, Sun C, et al. Object detection algorithm applied to optical genetic laser projection system[J]. Laser & Optoelectronics Progress, 2020, 57(6): 061503.
史再峰, 叶鹏, 孙诚, 等. 一种应用于光遗传激光投影系统的目标检测算法[J]. 激光与光电子学进展, 2020, 57(6): 061503.

[18] Zhou D W, Qiao X J, Zhang L J, et al. Parameters optimization of laser welding process of galvanized steel and 6016 aluminum alloy based on BP neural network and its microstructure and mechanical properties[J]. The Chinese Journal of Nonferrous Metals, 2014, 24(3): 678-688.
周惦武, 乔小杰, 张丽娟, 等. 镀锌钢/6016铝合金激光焊的BP神经网络工艺优化及组织和性能[J]. 中国有色金属学报, 2014, 24(3): 678-688.

[19] Nicolosi L, Tetzlaff R, Abt F, et al. Cellular Neural Network (CNN) based control algorithms for omnidirectional laser welding processes: experimental results[C]//2010 12th International Workshop on Cellular Nanoscale Networks and their Applications (CNNA 2010), February 3-5, 2010, Berkeley, CA, USA. New York: IEEE Press, 2010: 1-6.

[20] Guo D X, Chen J M, Cheng Y H. Laser cutting parameters optimization based on artificial neural network[C]//The 2006 IEEE International Joint Conference on Neural Network Proceedings, July 16-21, 2006, Vancouver, BC, Canada. New York: IEEE Press, 2006: 1106-1111.

[21] Sze V, Chen Y H, Yang T J, et al. Efficient processing of deep neural networks: a tutorial and survey[J]. Proceedings of the IEEE, 2017, 105(12): 2295-2329.