

基于多核处理器的多层感知神经网络设计和实现

朱新召¹, 胡哲琨², 周 莉¹, 陈 杰¹, 戈志伟¹

(1 中国科学院 微电子研究所, 北京 100029; 2 中国船舶重工集团第 709 所, 湖北 武汉 430079)

摘 要: 采用基于片上网络通信的多核处理器, 设计并实现了一种具有可重构性、运行时配置和更新, 以及具有学习能力的多层感知器神经网络结构以及对应的编程模型, 并在 FPGA 上进行了验证. 实验结果表明所设计的结构满足精度和实时性要求, 且编程模型简单, 易于扩展和维护, 为人工神经网络在计算机视觉和人工智能领域的应用提供了良好的解决方法.

关键词: 多层感知器; 神经网络; 多核处理器; 片上系统; FPGA

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2014)11-0027-05

The Design and Implementation of Reconfigurable Multilayer Perceptrons Neural Network Based on MPSoC

ZHU Xin-zhao¹, HU Zhe-kun², ZHOU Li¹, CHEN Jie¹, GE Zhi-wei¹

(1 Institute of Microelectronic, Chinese Academy of Sciences, Beijing 100029, China;

2 The 709th Research Institute of China Shipbuilding Industry Corporation, Wuhan 430079, China)

Abstract: A design and implementation of multilayer perceptrons based on an architecture of multiprocessor system-on-chip is proposed in this paper. The design is highly reconfigurable, can be configured and updated on runing-time, and has the property of learning. The design proposed in this paper is verified on FPGA. The result shows that the design meets the accuracy and real-time requirements, has simple programing interface, and is easy to extend and maintain, which would benefit the applications of neural networks in computer vision and artificial intelligence.

Key words: multilayer perceptrons; nernal network; multi processors; system on chip; FPGA

1 引言

人工神经网络(Artificial Neural Network, ANN)在模式分类、物体识别、数据统计^[1]、机器学习^[2]和信号处理领域有着广泛的应用. 人工神经网络的实现主要有软件和硬件两种方式. 采用软件的方式实现已经得到了很大的优化和改进, 然而其串行的特点导致性能不佳, 无法应用于实时性要求高的场合. 人工神经网络的模型可以并行计算, 这使得 ANN 非常适合硬件实现^[3]. 早从 1990 年开始已经有硬件实现人工神经网络的设计. 然而硬件实现面临的主要问题是资源的使用和互联的方案.

片上网络(Network on Chip, NoC)使用通用的网络连接代替了传统的定制互联策略, 因而使得系统具有更好的灵活性和伸缩性. 并且出现了一些基于 NoC 的片上多核处理器系统^[3](Multiprocessor Systems on Chip, MPSoC).

基于环面拓扑(torus topology)结构和固定路由策略, 实现了一个具有两个隐藏层的神经网络结构的仿真, 且系统具有低延迟和高吞吐率^[4]. 基于片上网络的思想^[5], 基于多核系统设计, 完成了神经网络在多核系统上的实现.

基于片上多核处理器系统平台, 本文设计并实现了可重构的多层感知器人工神经网络结构, 并且

收稿日期: 2014-01-03; **修回日期:** 2014-02-26

基金项目: 国家自然科学基金项目(61234003, 61036004, 61221004)

开发对应的编程模型,最终实现了系统的仿真和 FPGA 上的验证.

2 多层感知神经网络

2.1 多层感知神经网络结构

多层感知神经网络 (Multilayer Perceptron Neural Network, MLPNN) 中,各个感知器(神经元)以一种模拟人脑神经元的方式,由一系列单元相互密集连接构成^[6]. 多层神经网络包括一个输入层、一个或者多个隐藏层以及一个输出层. 图 1 即表示一个 6-8-2 的 MLP,包括一个输入层,一个隐藏层和一个输出层,其中输入层有 6 个输入单元,隐藏层有 8 个单元,而输出层有 2 个输出单元. 图中灰色部分为多层神经网络各层的偏置单元.

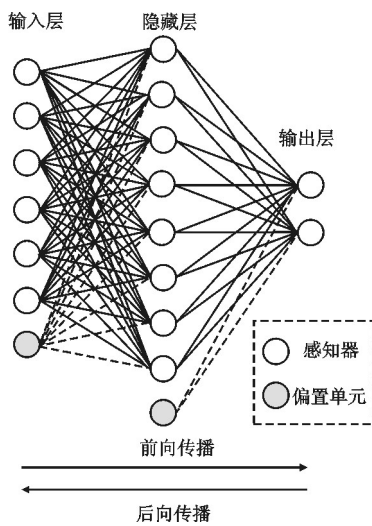


图 1 一个 6-8-2 的多层感知器网络

多层感知神经网络主要包括学习 (learning) 和推理 (inference) 两个阶段. 在学习阶段,根据训练的输入和目标向量,采用合适的算法拟合神经网络中各层之间的传递参数. 在推理阶段,输入数据按照前向的方式,经过多层感知神经网络各层,得到最终结果^[6].

2.2 多层感知器神经网络的后向传播算法

在多层神经网络的学习中广泛使用后向传播算法 (Back Propagation, BP)^[7]. BP 算法是一种有监督的学习算法,主要分为前向传播、误差后向传播和权值更新三个阶段. BP 算法每次接受一个模式输入,并据此更新神经网络中的权值.

2.2.1 前向传播阶段

将矢量 y_j 输入到神经网络中,并将信号沿 MLP 前向传播,计算出最终输出,如下式所示,其中

i 表示感知器编号,从 1 到 N_l , l 表示层编号,从 1 到总层数 L :

$$y_i^l = \varphi(u_i^l) \quad (1)$$

$$u_i^l = \sum_{j=0}^{N_{l-1}} w_{ij}^l y_j^{l-1} \quad (2)$$

$$1 \leq i \leq N_l, 1 \leq l \leq L \quad (3)$$

式中, w_{ij}^l 是 $(l-1)$ 层感知器 i 到 l 感知器 j 的权值,权值由训练过程获得; $\varphi(\cdot)$ 为激活函数,一般使用 sigmoid 函数或者 tanh 函数.

2.2.2 误差后向传播阶段

从输出层开始,误差按照反向传播的方式计算各层误差:

$$\delta_i^l = \varphi'(u_i^l)(t_i - y_i) \quad (4)$$

$$\delta_i^{l-1} = \varphi'(u_i^{l-1}) \sum_{i=1}^{N_l} w_{ij}^l \delta_i^l \quad (5)$$

$$1 \leq i \leq N_l, 1 \leq l \leq L \quad (6)$$

式中, t 为目标值, δ 是误差项,而 $\varphi'(\cdot)$ 是激活函数的一阶导数.

2.2.3 权值更新阶段

权值和偏置项更新如下:

$$\Delta w_{ij}^l(m) = \eta \delta_i^l y_j^{l-1} \quad (7)$$

$$w_{ij}^l(m) = w_{ij}^l(m-1) + \Delta w_{ij}^l(m) \quad (8)$$

$$1 \leq i \leq N_l, 1 \leq l \leq L \quad (9)$$

式中, Δw 是权值的改变量, η 为学习速率.

3 基于 MPSoC 的 MLP 设计

多层感知神经网络本身所具有的并行计算的特点,使之非常适合在多核处理器上实现. 片上多核处理器是基于一定的互联策略,在单个芯片上集成多个处理器核,以高度的并行性提升系统的性能. 本文所采用的 MPSoC 基于 NoC 互联策略,实现了多核之间数据的交互.

3.1 MPSoC 架构

如图 2 所示,本文所采用的基于 NoC 的片上多核处理器^[8]是基于网格的设计方法. 在 4×4 的 2D Mesh 片上网络中,集成 16 个节点,每个节点均为一个基于 AMBA 2.0 总线协议的小型系统. 节点系统包含一个小型 RISC 处理器、SRAM 控制器、大小可配置的 SRAM、中断控制器、片上网络路由器以及网络接口. 此外,节点 N(2, 2)中还包含了一个 UART 的控制模块,用于调试信息和结果的输出. 各节点使用定制的协议通信,包括数据交互和同步操作.

3.2 MLP 神经网络的实现

MLP 神经网络的 BP 算法,可以分为前向传播

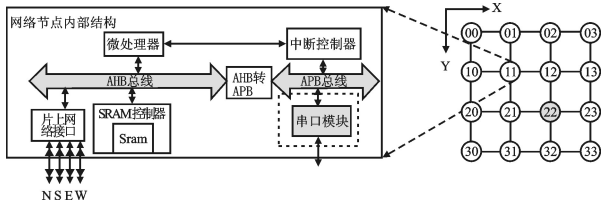


图2 多核处理器系统架构

和后向传播过程,其中前向传播过程包括BP算法的前向阶段,而后向传播过程包括BP算法中的误差后向传播和权值更新两个阶段。前向传播和后向传播可以简化为在相邻两层神经网络之间传递数据的过程。在MPSoC上实现时,每个物理核承载不同数据的虚拟感知器。MLP神经网络相邻层之间是全连接的,而在MPSoC中,并不存在这样的连接关系。

在设计的时候,采用如下策略将多层感知器神经网络映射到多核处理器上:前向传播的源层(即多层感知器神经网络前向传播中,靠近输入一侧的感知器层)感知器被均匀分配到各个物理核上;目的层(靠近输出一侧的感知器层)的感知器结果则同样分布在各个物理核上;每个分配了源层感知器的物理核同样分配这些感知器到所有目的层感知器的权值项;分配一个特殊的物理核M,同样保存所有目的层感知器的偏置项以及目的层误差项,物理核M被用来做归并结果产生、应用激活函数前的数据分发以及后向传播过程中误差项分发工作。

3.2.1 前向传播过程

MLP神经网络在MPSoC上的前向传播映射分为以下几步:

第一步,产生部分和。将源层感知器的输入 y_i^{l-1} 以及对应的权值集合 w_{ij}^l 分配到每一个处理核上,每个分配了源层感知器的处理核单独计算获得 $\sum_{i=1}^{N_l} w_{ij}^l y_j^{l-1}$ 的部分结果。

第二步,归并部分和。在这一步,每个处理核将其计算得到的部分和结果统一传送到处理核M上。所有数据接收完毕以后,处理核M将各个处理核的部分和累加获取 $\sum_{i=1}^{N_l} w_{ij}^l y_j^{l-1}$ 的结果。

第三步,加入偏置项。处理核M将目的层感知器的偏置项加入 $\sum_{i=1}^{N_l} w_{ij}^l y_j^{l-1}$,获得 u_i^l 的值。

第四步,分配并激活。激活函数一般由微处理核实现,且耗时较长。处理核M将上一步的结果分发到各个微处理核上,之后通过激活函数激活以获取目的层感知器的输出结果。

前向传播的流程图如图3所示,其中省略了一些同步过程:

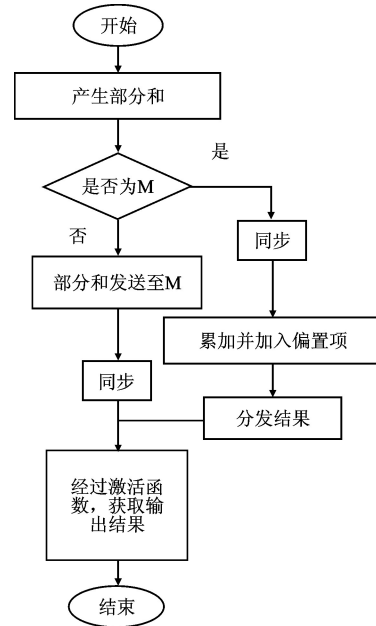


图3 前向传播流程图

3.2.2 后向传播过程

在反向传播过程中,对于输出层,其误差项由输出的理想结果直接和输出层结果做差获得;对于非输出层,其误差项则由后一层的误差按照式(5)反向传播到本层获得。误差项均保存在处理核M上。MLP在MPSoC上的后向传播部分主要由以下几步:

第一步,分发目的层的误差项。处理核M将其保存的目的层误差项发送到各个源层所在的物理核上。

第二步,计算局部梯度值。根据式(7),各物理核由保存的源层数据以及源层到目的层的权值,计算得到局部梯度值。

第三步,计算源层误差。根据式(5),各物理核计算源层的误差项。

第四步,更新权值和偏置项。根据式(8),更新各处理核保存的权值。对于处理核M,还需要更新其保存的偏置项。

第五步,归并源层误差。各个处理核将其计算得到的源层误差项发送到物理核M上,物理核M归并源层误差项,为下一次反向传播准备数据。

后向传播的流程图如图4所示,其中省略了一些同步过程:

3.2.3 可重构设计

基于多核处理器系统的MLP实现是可重构

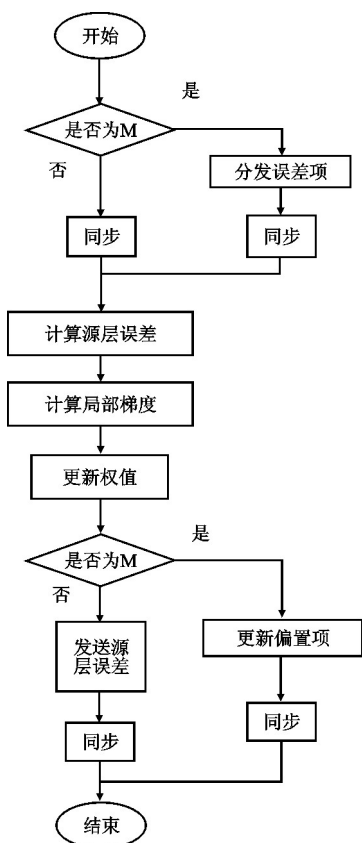


图4 后向传播流程图

的,这主要表现在以下几个方面:

(1)MLP 的结构配置可以修改,因而可以使用同一个 MPSoC 实现不同结构的多层感知器神经网络。

(2)MLP 的权值存储在 SRAM 中,可以很容易通过编程的方式更改其权值。

(3)神经网络的激活函数可以根据需要配置为不同的函数。

3.3 数据结构和编程接口

本文设计了保存每个物理核中的感知器单元以及权值集合的数据结构,并实现了物理核内部的多层感知器神经网络计算,如图5所示。物理核之间的通信由 MPSoC 提供的相关函数实现。

MATRIX_MULT_NODE 用于前向传播过程中,由输入值和权值生成输出值的部分和结果,是由函数 matrix_mult() 完成的。而 WEIGHT_DELTA_NODE 则用于后向传播过程中,由获取的局部梯度值和学习速率更新权值,由函数 adjust_weight() 完成。

4 系统仿真和性能分析

4.1 仿真平台和配置

本文以参考文献[6]中的人脸识别的多层感知

```

//每个物理核存储的输入值,权值
//以及输出的部分和
typedef struct{
    ANN_DATATYPE*in;//输入数据指针
    ANN_DATATYPE*w;//权值指针
    ANN_DATATYPE*out;//部分输出指针
    uint16_t in_n;//输入数据个数
    uint16_t out_n;//输出数据个数
}MATRIX_MULT_NODE;
//每个物理核用于调整权值的数据结构
typedef struct{
    ANN_DATATYPE*in;//输入数据指针
    ANN_DATATYPE*w;//权值指针
    ANN_DATATYPE*delta;//局部梯度指针
    uint16_t in_n;//输入数据个数
    uint16_t out_n;//输出数据个数
}WEIGHT_DELTA_NODE;

//前向传播过程
//根据输入以及权值计算部分输出值
void matrix_mult(MATRIX_MULT_NODE*P);

//调整权值部分
//根据获取的局部梯度以及学习速率调整权值
void adjust_weight(WEIGHT_DELTA_NODE*P,
const float eta);
  
```

图5 数据结构和编程接口

器神经网络为例,在 Altera 公司推出的 Stratix II EP2S180 FPGA 开发板上进行了仿真验证。

表1所示为本系统和对比系统的配置情况,其中 N.A 代表“无法获取数据”(Not Avilable),即参考文献中未提及关于这些项目的信息。“精度”一项是结合感知器表示数据的精度以及权值精度的结果,例如“32b×32b”表示感知器内的数据和权值均使用单精度浮点型数据。

表1 系统和 MLP 配置

平台	系统结构	MLP 拓扑	精度	学习方式	硬件浮点运算支持
参考文献[6]	单核	20-3-1	32 b×32 b	离线	N.A
参考文献[7]	2×3 网络	(100,100)×6	32 b×32 b	离线	可配置
本文	4×4 网络	960-4-1	32 b×32 b	在线	否

本文所提出的系统采用 4×4 网络结构,感知器表示数据和权值都采用单精度浮点型格式,不支持硬件浮点加速,浮点运算使用软件模拟的方式进行,系统支持在线的学习方式,即学习过程可以在系统上完成。而参考文献[5]系统的学习过程是使用离线的方式完成的。

4.2 性能比较

评价硬件神经网络的性能的主要参数^[8]是每秒钟连接数目(Connection Per Second, CPS),即神经网络每秒钟执行乘累加的数目。对于多核系统,每核每秒钟连接数目(Connection Per Second Per Core, CPSPC)也是一个非常重要的评价指标,CPSPC=CPS/N_c,其中 N_c是硬件中的核数。对于具有在线学习能力的神经网络而言,每秒钟更新数目(Con-

nection Update Per Second, CUPS)定义为神经网络每秒钟内可以更新的权值和偏置项的数目。

未使用浮点加速时,参考文献[5]中, 2×3 的结构前向传播的发射频率(Shooting Frequency)为 145 Hz,连接的总数目为 60 000,故其每个核的 CPS 为 1 450 000。

各系统性能对比见表 2。其中,参考文献[5]中的 NoC 平台不具备在线学习的功能,因此其 CUPS 项的数据无效。图 6 为归一化的性能比较图,其中以本系统的性能表现为基准,将其他两个系统的性能进行了归一化表示。

表 2 MPSoC 实现 MLP 性能对比

平台	总连接数	CPS	CPSPC	CUPS
参考文献[6]	67	115 000	115 000	N. A
参考文献[7]	60 000	8 700 000	1 450 000	N. A
本文	3 849	27 069 840	1 691 865	42 949 672

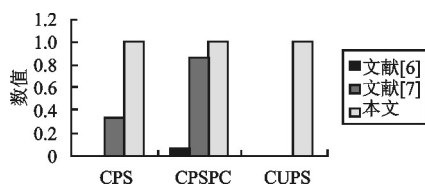


图 6 归一化性能比较

从图 6 中可以看出,未使用硬件浮点加速时,相比于其他两个系统,本系统具有最高的 CPS 值,以及 CPSPC 值,表明本系统性能优于其他两个系统。

5 结束语

基于多核处理器结构和并行编程模型的思想,本文设计并实现了多层感知神经网络。本文所设计的基于多核处理器的多层感知神经网络具有良好的可重构性,编程模型简单,易于维护和扩展。实验结果表明,本文设计的多层感知神经网络具有良

好的性能表现,满足实时性要求,为神经网络在其他

领域的使用提供了良好的解决方法。

参考文献:

- [1] Bengio Y. Learning deep architectures for AI[J]. Foundations and trends in Machine Learning, 2009, 2 (1): 1-127.
- [2] GironÉS R G, Palero R C, Boluda J C, et al. FPGA implementation of a pipelined on-line backpropagation [J]. Journal of VLSI signal processing systems for signal, image and video technology, 2005, 40 (2): 189-213.
- [3] Yiping D, Takahiro W. High performance NoC architecture for two hidden layers BP Neural Network[J]. IEEE, 2008(1):269-272.
- [4] Men shen T, Yao hsien F. Implementation of high performance hardware based toroidal neural network with learning capability[C]//proceeding of the Advanced Intelligent Mechatronics 2009, Singapore, 2009: 180-185.
- [5] Mand N P, Robino F, Oberg J. Artificial neural network emulation on NOC based multi-core FPGA platform[J]. Norchip, 2012(4):4-7.
- [6] Mitcheli T. 机器学习[M]. 曾华军, 张银奎, 译. 北京:机械工业出版社, 2008.
- [7] Duda R O, Hart P E, Stork D G. 模式分类[M]. 李宏东, 译. 北京:机械工业出版社, 2003.
- [8] Dias F M, Antunes A, Mota A M. Artificial neural networks: a review of commercial hardware[J]. Engineering Applications of Artificial Intelligence, 2004, 17(8): 945-952.

作者简介:

朱新召 男, (1990-), 硕士研究生. 研究方向为集成电路工程。

(上接第 26 页)

- [9] 周宣武. 具有附加性质的签密方案及其应用研究[D]. 天津:天津大学, 2013.

作者简介:

周宣武 男, (1980-), 博士后. 研究方向为密码学与信息安全。

付 燕 女, (1985-), 博士研究生. 研究方向为可信计算。

金志刚 男, (1973-), 教授, 博士生导师. 研究方向为计算机网络与信息安全。

刘开华 男, (1956-), 教授, 博士生导师. 研究方向为计算机网络、软件无线电技术。

李 欣 男, (1965-), 副教授. 研究方向为网络安全。