(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 107545303 B
(45)授权公告日 2021.09.07

(21)申请号 201710794580.X
(22)申请日 2016.01.20
(65)同一申请的已公布的文献号
申请公布号 CN 107545303 A
(43)申请公布日 2018.01.05
(62)分案原申请数据
201610039162.5 2016.01.20
(73)专利权人 中科寒武纪科技股份有限公司
地址 100190 北京市海淀区科学院南路6号科研综合楼644室
(72)发明人 不公告发明人
(74)专利代理机构 中科专利商标代理有限责任公司 11021
代理人 任岩
(51)Int.Cl.
G06N 3/04(2006.01)
G06N 3/06(2006.01)
审查员 江汉琼

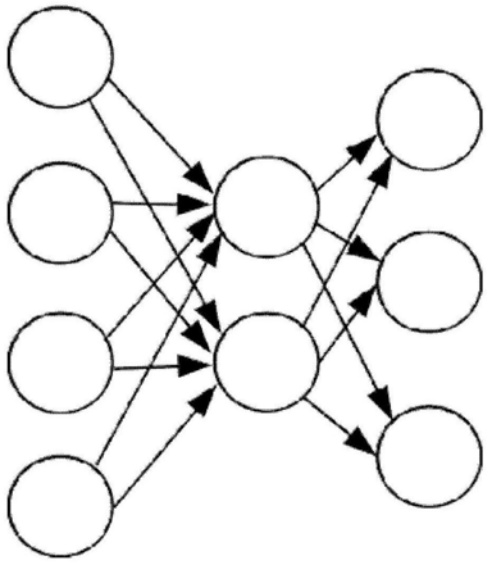
权利要求书4页 说明书11页 附图9页

(54)发明名称

用于稀疏人工神经网络的计算装置和运算方法

(57)摘要

一种计算装置和方法，所述计算装置包括运算单元，用于接收指令对所述权值和所述相应的输入神经元执行用于稀疏连接的人工神经网络运算，得到输出神经元。本公开的装置解决了CPU和GPU运算性能不足，前端译码开销大的问题，有效提高了对人工神经网络运算算法的支持，避免了内存带宽成为人工神经网络运算及其训练算法性能瓶颈的问题。



CN 107545303 B

1.一种用于稀疏人工神经网络的计算装置,其特征在于,包括:

映射单元,用于接收输入数据,输入数据包括输入神经元、连接关系和权值;然后根据连接关系将每个权值一一对应于相应的输入神经元,得到处理后的神经元和处理后的权值;其中,所述连接关系表示每个输入神经元和每个输出神经元是否有对应的连接关系的权值;

运算单元,用于接收指令、处理后的神经元和处理后的权值,然后根据指令对所述处理后的神经元和处理后的权值执行人工神经网络运算,得到输出神经元。

2.如权利要求1所述的用于稀疏人工神经网络的计算装置,其特征在于,所述用于稀疏人工神经网络的计算装置还包括存储装置,用于存储输入数据、权值数据、连接关系数据和指令;所述连接关系包括:

第一种情形:

采用1表示有连接,0表示无连接,每个输出神经元与所有输入输出神经元的连接状态组成一个0和1的字符串来表示该输出神经元的连接关系;或者

采用1表示有连接,0表示无连接,每个输入神经元与所有输出神经元的连接状态组成一个0和1的字符串来表示该输入神经元的连接关系;或者

第二种情形:

将一输出神经元的第一个有连接的输入神经元所在的位置距离第一个输入神经元的距离、所述输出神经元的第二个有连接的输入神经元距离第一个有连接的输入神经元的距离,所述输出神经元的第三个有连接的输入神经元距离第二个有连接的输入神经元距离,……,依次类推,直到穷举所述输出神经元的所有输入神经元,来表示所述输出的连接关系。

3.如权利要求1或2所述的用于稀疏人工神经网络的计算装置,其特征在于,所述运算单元包括乘法器、加法树和非线性函数单元;所述运算单元接收指令对所述处理后的神经元和处理后的权值执行运算,得到输出神经元包括:所述乘法器将所述处理后的神经元和处理后的权值相乘,得到加权神经元数据;和/或,所述加法树将加权神经元数据相加得到总加权神经元数据;和/或,所述非线性函数单元对总加权神经元数据执行激活函数运算,得到输出神经元。

4.如权利要求3所述的用于稀疏人工神经网络的计算装置,其特征在于,所述运算单元将加权神经元数据相加得到总加权神经元数据包括将加权神经元数据执行加法树运算得到总加权神经元数据。

5.如权利要求4所述的用于稀疏人工神经网络的计算装置,其特征在于,所述运算单元将总加权神经元数据和偏置相加得到加偏置神经元数据,和/或,对加偏置神经元数据执行激活函数运算,得到输出神经元。

6.如权利要求5所述的用于稀疏人工神经网络的计算装置,其特征在于,所述激活函数包括sigmoid函数、tanh函数或ReLU函数。

7.如权利要求6所述的用于稀疏人工神经网络的计算装置,其特征在于,所述用于稀疏人工神经网络的计算装置包括存储装置,所述存储装置还用于存储输出神经元。

8.如权利要求7所述的用于稀疏人工神经网络的计算装置,其特征在于,映射单元将部分或全部的相应的输入神经元存储在存储装置中。

9.如权利要求8所述的用于稀疏人工神经网络的计算装置,其特征在于,所述用于稀疏人工神经网络的计算装置还包括:

指令缓存,用于存储所述指令;以及

控制单元,用于从所述指令缓存中读取指令,并将读取的指令译码。

10.如权利要求9所述的用于稀疏人工神经网络的计算装置,其特征在于,所述用于稀疏人工神经网络的计算装置还包括:输入神经元缓存,用于缓存输入神经元;和权值缓存,用于缓存权值。

11.如权利要求10所述的用于稀疏人工神经网络的计算装置,其特征在于,所述用于稀疏人工神经网络的计算装置还包括:输出神经元缓存,用于缓存输出神经元。

12.如权利要求11所述的用于稀疏人工神经网络的计算装置,其特征在于,所述用于稀疏人工神经网络的计算装置还包括直接内存存取单元,用于在所述存储装置、指令缓存、输入神经元缓存、输出神经元缓存和权值缓存中进行数据或指令读写。

13.如权利要求10所述的用于稀疏人工神经网络的计算装置,其特征在于,映射单元将所述相应的输入神经元存储在输入神经元缓存中。

14.如权利要求1、2、4-12中任一项所述的用于稀疏人工神经网络的计算装置,其特征在于,映射单元将相应的输入神经元直接传输给运算单元。

15.如权利要求1、2、4-13中任一项所述的用于稀疏人工神经网络的计算装置,其特征在于,所述运算单元采用针对稀疏的多层人工神经网络运算的专用SIMD指令。

16.如权利要求3所述的用于稀疏人工神经网络的计算装置,其特征在于,所述运算单元采用针对稀疏的多层人工神经网络运算的专用SIMD指令。

17.如权利要求14所述的用于稀疏人工神经网络的计算装置,其特征在于,所述运算单元采用针对稀疏的多层人工神经网络运算的专用SIMD指令。

18.一种包括如权利要求1-17中任一所述的用于稀疏人工神经网络的计算装置的用于稀疏人工神经网络的运算装置,其特征在于,还包括CPU或GPU,用于数据搬运,控制所述用于稀疏人工神经网络的计算装置,传输指令给所述用于稀疏人工神经网络的计算装置;以及I/O接口,用于处理CPU或GPU的输入/输出;所述用于稀疏人工神经网络的运算装置用于接受CPU或GPU的数据和指令,执行稀疏的多层人工神经网络运算算法得到执行结果,然后将执行结果传输给CPU或GPU;所述用于稀疏人工神经网络的运算装置包括存储单元,还用于存储稀疏的多层人工神经网络模型。

19.如权利要求18所述的用于稀疏人工神经网络的运算装置,其特征在于,包括多个所述用于稀疏人工神经网络的计算装置。

20.如权利要求19所述的用于稀疏人工神经网络的运算装置,其特征在于,多个所述用于稀疏人工神经网络的计算装置通过PCIE总线互联,多个所述用于稀疏人工神经网络的计算装置共用同一个宿主CPU或GPU或者分别有相应的宿主CPU或GPU。

21.如权利要求20所述的用于稀疏人工神经网络的运算装置,其特征在于,多个所述用于稀疏人工神经网络的计算装置共享指令缓存、输入神经元缓存、输出神经元缓存和/或权值缓存。

22.一种用于稀疏人工神经网络的运算方法,其特征在于,所述方法包括:

使用映射单元接收输入数据和存储单元存储的权值和连接关系数据,输入数据包括输

入神经元、连接关系和权值;然后使用映射单元根据连接关系数据将每个权值一一对应于相应的输入神经元,得到处理后的神经元和处理后的权值;其中,所述连接关系数据为表示每个输入神经元和每个输出神经元是否有对应的连接关系的权值;以及

使用运算单元接收指令对所述处理后的神经元和处理后的权值执行运算,得到输出神经元。

23.如权利要求22所述的用于稀疏人工神经网络的运算方法,其特征在于,所述连接关系包括:

第一种情形:

采用1表示有连接,0表示无连接,每个输出神经元与所有输入输出神经元的连接状态组成一个0和1的字符串来表示该输出神经元的连接关系;或者

采用1表示有连接,0表示无连接,每个输入神经元与所有输出神经元的连接状态组成一个0和1的字符串来表示该输入神经元的连接关系;或者

第二种情形:

将一输出神经元的第一个有连接的输入神经元所在的位置距离第一个输入神经元的距离、所述输出神经元的第二个有连接的输入神经元距离第一个有连接的输入神经元的距离,所述输出神经元的第三个有连接的输入神经元距离第二个有连接的输入神经元距离,……,依次类推,直到穷举所述输出神经元的所有输入神经元,来表示所述输出的连接关系。

24.如权利要求22或23所述的用于稀疏人工神经网络的运算方法,其特征在于,所述运算单元接收指令对所述处理后的神经元和处理后的权值执行运算,得到输出神经元的步骤包括:所述运算单元将所述处理后的神经元和处理后的权值相乘,得到加权神经元数据;将加权神经元数据相加得到总加权神经元数据;和/或,对总加权神经元数据执行激活函数运算,得到输出神经元。

25.如权利要求24所述的用于稀疏人工神经网络的运算方法,其特征在于,所述运算单元将加权神经元数据相加得到总加权神经元数据的步骤包括:将加权神经元数据执行加法树运算得到总加权神经元数据。

26.如权利要求25所述的用于稀疏人工神经网络的运算方法,其特征在于,所述运算单元将总加权神经元数据和偏置相加得到加偏置神经元数据,和/或,对加偏置神经元数据执行激活函数运算,得到输出神经元。

27.如权利要求26所述的用于稀疏人工神经网络的运算方法,其特征在于,所述激活函数包括sigmoid函数、tanh函数或ReLU函数。

28.如权利要求27所述的用于稀疏人工神经网络的运算方法,还包括使用存储装置暂存储输出神经元。

29.如权利要求28所述的用于稀疏人工神经网络的运算方法,其特征在于,映射单元将部分或全部的相应的输入神经元存储在存储装置中。

30.如权利要求22、23、25-29中任一项所述的用于稀疏人工神经网络的运算方法,其特征在于,所述用于稀疏人工神经网络的运算方法还包括:

使用指令缓存存储所述指令;以及

使用控制单元从所述指令缓存中读取指令,并将读取的指令译码。

31.如权利要求30所述的用于稀疏人工神经网络的运算方法，其特征在于，所述用于稀疏人工神经网络的运算方法还包括：使用输入神经元缓存缓存输入神经元；以及使用权值缓存缓存权值。

32.如权利要求31所述的用于稀疏人工神经网络的运算方法，其特征在于，所述用于稀疏人工神经网络的运算方法还包括：使用输出神经元缓存缓存输出神经元。

33.如权利要求32所述的用于稀疏人工神经网络的运算方法，其特征在于，所述用于稀疏人工神经网络的运算方法还包括使用直接内存存取单元在所述存储装置、指令缓存、输入神经元缓存、输出神经元缓存和权值缓存中进行数据或指令读写。

34.如权利要求31所述的用于稀疏连接的人工神经网络运算方法，其特征在于，映射单元将相应的输入神经元存储在输入神经元缓存中。

35.如权利要求22、23、25-29、32、33中任一项所述的用于稀疏连接的人工神经网络运算方法，其特征在于，映射单元将相应的输入神经元直接传输给运算单元。

36.如权利要求22、23、25-29、31-34中任一项所述的用于稀疏人工神经网络的运算方法，其特征在于，所述用于稀疏人工神经网络的运算方法采用针对稀疏的多层人工神经网络运算的专用SIMD指令。

37.如权利要求24所述的用于稀疏人工神经网络的运算方法，其特征在于，所述用于稀疏人工神经网络的运算方法采用针对稀疏的多层人工神经网络运算的专用SIMD指令。

38.如权利要求30所述的用于稀疏人工神经网络的运算方法，其特征在于，所述用于稀疏人工神经网络的运算方法采用针对稀疏的多层人工神经网络运算的专用SIMD指令。

39.如权利要求35所述的用于稀疏人工神经网络的运算方法，其特征在于，所述用于稀疏人工神经网络的运算方法采用针对稀疏的多层人工神经网络运算的专用SIMD指令。

40.一种包括如权利要求36-39任一项所述的用于稀疏人工神经网络的运算方法的稀疏的多层人工神经网络运算方法，其特征在于，还包括使用CPU或GPU执行数据搬运，控制所述用于稀疏人工神经网络的计算装置，传输指令给所述用于稀疏人工神经网络的计算装置；以及使用I/O接口处理CPU或GPU的输入/输出；所述用于稀疏人工神经网络的运算装置接受CPU或GPU的数据和指令，执行稀疏的多层人工神经网络运算算法得到执行结果，然后将执行结果传输给CPU或GPU；所述用于稀疏人工神经网络的运算装置包括存储单元，用于存储稀疏的多层人工神经网络模型。

41.如权利要求40所述的稀疏的多层人工神经网络运算方法，其特征在于，所述稀疏的多层人工神经网络运算方法中包括并行执行多个所述人工神经网络运算方法的步骤。

用于稀疏人工神经网络的计算装置和运算方法

[0001] 本公开是申请日为2016年1月20日、申请号为201610039162.5、发明名称为“一种用于稀疏连接的人工神经网络计算装置和方法”的发明专利申请的分案申请。

技术领域

[0002] 本公开涉及数据处理技术领域，更具体地涉及一种用于稀疏人工神经网络的计算装置和运算方法。

背景技术

[0003] 人工神经网络(Artificial Neural Networks，ANNs)简称为神经网络(NNs)，它是一种模仿动物神经网络行为特征，进行分布式并行信息处理的算法数学模型。这种网络依靠系统的复杂程度，通过调整内部大量节点之间的相互连接关系，从而达到处理信息的目的。神经网络用到的算法就是向量乘法，并且广泛采用符号函数及其各种逼近。

[0004] 就像大脑里的神经网络一样，神经网络由一些互相连接的节点组成，如图1所示，每个圆圈表示一个神经元，每个箭头表示两个神经元之间的连接又被称为权值。

[0005] 神经元的计算公式可以简单的描述成：$y=f(\sum_{i=0}^{n} w_i * x_i)$。其中，x表示所有和输出神经元相连接的输入神经元，w表示x和输出神经元之间对应的权值。f(x)是一个非线性函数，通常称作激活函数，常用的函数如：$\frac{1}{1+e^{-x}}$，$\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$等。

[0006] 神经网络被广泛应用于各种应用场景：计算视觉、语音识别和自然语言处理等。在近几年的时间里，神经网络的规模一直在增长。在1998年，Lecun用于手写字符识别的神经网络的规模小于1M个权值；在2012年，krizhevsky用于参加ImageNet竞赛的规模是60M个权值。

[0007] 神经网络是一个高计算量和高访存的应用，权值越多，计算量和访存量都会增大。为了减小计算量和权值数量，从而降低访存量，出现了稀疏连接的神经网络，如图2所示即为一个稀疏的神经网络。

[0008] 随着神经网络计算量和访存量的急剧增大，现有技术中通常采用通用处理器计算稀疏的人工神经网络。对于通用处理器，输入神经元、输出神经元和权值分别存储在三个数组中，同时还有一个索引数组，索引数组存储了每个输出神经元和输入神经元通过权值连接的连接关系。在计算时，主要的运算是神经元与权值相乘。每一次运算都要通过索引数组找到神经元对应的权值。由于通用处理器计算能力和访存能力都很弱，满足不了神经网络的需求。而多个通用处理器并行执行时，通用处理器之间相互通讯又成为了性能瓶颈。在计算剪枝之后的神经网络时，每次乘法运算都要去索引数组里重新查找权值对应的位置，增加了额外的计算量和访存开销。因此计算神经网络耗时长，功耗高。通用处理器需要把多层人工神经网络运算译码成一长列运算及访存指令序列，处理器前端译码带来了较大的功耗开销。

[0009] 另一种支持稀疏连接的人工神经网络运算及其训练算法的已知方法是使用图形

处理器(GPU),该方法通过使用通用寄存器堆和通用流处理单元执行通用SIMD指令来支持上述算法。但由于GPU是专门用来执行图形图像运算以及科学计算的设备,没有对稀疏的人工神经网络运算的专门支持,仍然需要大量的前端译码工作才能执行稀疏的人工神经网络运算,带来了大量的额外开销。另外GPU只有较小的片上缓存,多层人工神经网络的模型数据(权值)需要反复从片外搬运,片外带宽成为了主要性能瓶颈,同时带来了巨大的功耗开销。

发明内容

[0010] 有鉴于此,本公开的目的在于提供一种用于稀疏人工神经网络的计算装置和运算方法。

[0011] 为了实现上述目的,作为本公开的一个方面,本公开提供了一种用于稀疏人工神经网络的计算装置,包括:

[0012] 存储装置,用于存储输入数据、一个或多个权值数据、连接关系数据和指令;所述连接关系数据表示每个输入神经元数据和每个输出神经元数据是否有对应的连接关系的权值数据;

[0013] 映射单元,用于接收输入数据,输入数据包括输入神经元;然后根据连接关系将每个权值一一对应于相应的输入神经元和权值;

[0014] 运算单元,用于接收指令对所述权值和所述相应的输入神经元执行人工神经网络运算,得到输出神经元。

[0015] 其中,所述连接关系包括:

[0016] 第一种情形:

[0017] 采用1表示有连接,0表示无连接,每个输出神经元与所有输入输出神经元的连接状态组成一个0和1的字符串来表示该输出神经元的连接关系;或者

[0018] 采用1表示有连接,0表示无连接,每个输入神经元与所有输出神经元的连接状态组成一个0和1的字符串来表示该输入神经元的连接关系;或者

[0019] 第二种情形:

[0020] 将一输出神经元的第一个有连接的输入神经元所在的位置距离第一个输入神经元的距离、所述输出神经元的第二个有连接的输入神经元距离第一个有连接的输入神经元的距离,所述输出神经元的第三个有连接的输入神经元距离第二个有连接的输入神经元距离,……,依次类推,直到穷举所述输出神经元的所有输入神经元,来表示所述输出的连接关系。

[0021] 其中,所述运算单元包括乘法器、加法树和非线性函数单元;所述运算单元接收指令对所述相应的输入神经元和所述权值执行运算,得到输出神经元包括:所述乘法器将所述相应的输入神经元和权值相乘,得到加权神经元数据;和/或,所述加法树将加权神经元数据相加得到总加权神经元数据;和/或所述非线性函数单元对总加权神经元数据执行激活函数运算,得到输出神经元。

[0022] 其中,所述运算单元将加权神经元数据相加得到总加权神经元数据包括将加权神经元数据执行加法树运算得到总加权神经元数据。

[0023] 其中,所述运算单元将总加权神经元数据和偏置相加得到加偏置神经元数据,和/

或对加偏置神经元数据执行激活函数运算，得到输出神经元。

[0024] 其中，所述激活函数包括sigmoid函数、tanh函数或ReLU函数。

[0025] 其中，所述存储装置还用于暂存储输出神经元。

[0026] 其中，映射单元将部分或全部的相应的输入神经元存储在存储装置中。

[0027] 其中，所述用于稀疏连接的人工神经网络计算装置还包括：

[0028] 指令缓存，用于存储所述指令；以及

[0029] 控制单元，用于从所述指令缓存中读取指令，并将读取的指令译码。

[0030] 其中，所述用于稀疏人工神经网络的计算装置还包括：输入神经元缓存，用于缓存输入神经元；以及权值缓存，用于缓存权值。

[0031] 其中，所述用于稀疏人工神经网络的计算装置还包括：输出神经元缓存，用于缓存输出神经元。

[0032] 其中，所述用于稀疏人工神经网络的计算装置还包括直接内存存取单元，用于在所述存储装置、指令缓存、输入神经元缓存、输出神经元缓存和权值缓存中进行数据或指令读写。

[0033] 其中，映射单元将所述相应的输入神经元存储在输入神经元缓存中。

[0034] 其中，映射单元将相应的输入神经元直接传输给运算单元。

[0035] 其中，所述运算单元采用针对稀疏的多层人工神经网络运算的专用SIMD指令。

[0036] 作为本公开的另一个方面，本公开还提供了一种包括如上所述的用于稀疏人工神经网络的计算装置的人工神经网络运算装置，还包括CPU或GPU，用于数据搬运，控制所述用于稀疏人工神经网络的计算装置，传输指令给所述用于稀疏人工神经网络的计算装置；以及I/O接口，用于处理CPU或GPU的输入/输出；所述运算装置用于接受CPU或GPU的数据和指令，执行稀疏的多层人工神经网络运算算法得到执行结果，然后将执行结果传输给CPU或GPU；所述用于稀疏人工神经网络的计算装置包括存储单元，还用于存储稀疏的多层人工神经网络模型。

[0037] 其中，所述人工神经网络运算装置包括多个所述用于稀疏人工神经网络的计算装置。

[0038] 其中，多个所述用于稀疏人工神经网络的计算装置通过PCIE总线互联，多个所述用于稀疏人工神经网络的计算装置共用同一个宿主CPU或GPU或者分别有相应的宿主CPU或GPU。

[0039] 其中，多个所述用于稀疏人工神经网络的计算装置共享指令缓存、输入神经元缓存、输出神经元缓存和/或权值缓存。

[0040] 作为本公开的再一个方面，本公开还提供了一种用于稀疏人工神经网络的运算方法，所述方法包括：

[0041] 使用映射单元接收输入数据和存储单元存储的权值和连接关系，输入数据包括输入神经元；然后根据连接关系将每个权值一一对应于相应的输入神经元和权值；以及

[0042] 使用运算单元接收指令对所述权值和所述相应的输入神经元执行运算，得到输出神经元。

[0043] 其中，所述连接关系包括：

[0044] 第一种情形：

[0045] 采用1表示有连接,0表示无连接,每个输出神经元与所有输入输出神经元的连接状态组成一个0和1的字符串来表示该输出神经元的连接关系;或者

[0046] 采用1表示有连接,0表示无连接,每个输入神经元与所有输出神经元的连接状态组成一个0和1的字符串来表示该输入神经元的连接关系;或者

[0047] 第二种情形:

[0048] 将一输出神经元的第一个有连接的输入神经元所在的位置距离第一个输入神经元的距离、所述输出神经元的第二个有连接的输入神经元距离第一个有连接的输入神经元的距离,所述输出神经元的第三个有连接的输入神经元距离第二个有连接的输入神经元距离,……,依次类推,直到穷举所述输出神经元的所有输入神经元,来表示所述输出的连接关系。

[0049] 其中,所述运算单元接收指令对所述相应的输入神经元和所述权值执行运算,得到输出神经元的步骤包括:所述运算单元将所述相应的输入神经元和权值相乘,得到加权神经元数据;将加权神经元数据相加得到总加权神经元数据;和/或,对总加权神经元数据执行激活函数运算,得到输出神经元。

[0050] 其中,所述运算单元将加权神经元数据相加得到总加权神经元数据的步骤包括:将加权神经元数据执行加法树运算得到总加权神经元数据。

[0051] 其中,所述运算单元将总加权神经元数据和偏置相加得到加偏置神经元数据,和/或,对加偏置神经元数据执行激活函数运算,得到输出神经元。

[0052] 其中,所述激活函数包括sigmoid函数、tanh函数或ReLU函数。

[0053] 其中,所述用于稀疏人工神经网络的运算方法还包括使用存储装置暂时存储输出神经元。

[0054] 其中,映射单元将部分或全部的相应的输入神经元存储在存储装置中。

[0055] 其中,所述用于稀疏人工神经网络的运算方法还包括:

[0056] 使用指令缓存存储所述指令;以及

[0057] 使用控制单元从所述指令缓存中读取指令,并将读取的指令译码。

[0058] 其中,所述用于稀疏人工神经网络的运算方法还包括:使用输入神经元缓存缓存输入神经元;以及使用权值缓存缓存权值。

[0059] 其中,所述用于稀疏人工神经网络的运算方法还包括:使用输出神经元缓存缓存输出神经元。

[0060] 其中,所述用于稀疏人工神经网络的运算方法还包括使用直接内存存取单元在所述存储装置、指令缓存、输入神经元缓存、输出神经元缓存和权值缓存中进行数据或指令读写。

[0061] 其中,映射单元将相应的输入神经元存储在输入神经元缓存中。

[0062] 其中,映射单元将相应的输入神经元直接传输给运算单元。

[0063] 其中,所述用于稀疏人工神经网络的运算方法采用针对稀疏的多层人工神经网络运算的专用SIMD指令。

[0064] 作为本公开的还一个方面,本公开还提供了一种包括如上所述的用于稀疏人工神经网络的运算方法的稀疏的多层人工神经网络运算方法,还包括使用CPU或GPU执行数据搬运,控制所述用于稀疏人工神经网络的计算装置,传输指令给所述用于稀疏人工神经网络

的计算装置;以及使用I/O接口处理CPU或GPU的输入/输出;所述用于稀疏人工神经网络的运算装置接受CPU或GPU的数据和指令,执行稀疏的多层人工神经网络运算算法得到执行结果,然后将执行结果传输给CPU或GPU;所述用于稀疏人工神经网络的运算装置包括存储单元,用于存储稀疏的多层人工神经网络模型。

[0065] 其中,所述稀疏的多层人工神经网络运算方法包括并行执行多个所述用于稀疏人工神经网络的运算方法的步骤。

[0066] 基于上述技术方案可知,本公开的用于稀疏人工神经网络的计算装置和运算方法至少具有以下有益效果之一:

[0067] (1)通过采用针对稀疏的多层人工神经网络运算的专用SIMD指令和定制的运算单元,解决了CPU和GPU运算性能不足,前端译码开销大的问题,有效提高了对多层人工神经网络运算算法的支持;

[0068] (2)通过采用针对多层人工神经网络运算算法的专用片上缓存,充分挖掘了输入神经元和权值数据的重用性,避免了反复向内存读取这些数据,降低了内存访问带宽,避免了内存带宽成为多层人工神经网络运算及其训练算法性能瓶颈的问题。

附图说明

[0069] 图1是神经网络的节点结构示意图;

[0070] 图2是稀疏连接的神经网络的节点结构示意图;

[0071] 图3是作为本公开一实施例的总体结构的示意性框图;

[0072] 图4是作为本公开一实施例的一稀疏连接的神经网络的节点结构示意图;

[0073] 图5是图4的神经网络的连接关系示意图;

[0074] 图6是作为本公开又一实施例的一稀疏连接的神经网络的连接关系示意图;

[0075] 图7是作为本公开一实施例的一卷积操作的示意图;

[0076] 图8是卷积神经网络变得稀疏时输入、输出和权值的变化图;

[0077] 图9是作为本公开一实施例的稀疏连接的人工神经网络运算装置的结构示意图;

[0078] 图10是作为本公开一实施例的映射单元的结构示意图;

[0079] 图11是作为本公开一实施例的稀疏连接的人工神经网络运算过程的流程图;

[0080] 图12是作为本公开另一实施例的稀疏连接的人工神经网络运算装置的结构示意图;

[0081] 图13是作为本公开另一实施例的映射单元的结构示意图;

[0082] 图14是作为本公开再一实施例的稀疏连接的人工神经网络运算装置的结构示意图;

[0083] 图15是作为本公开再一实施例的映射单元的结构示意图;

[0084] 图16是作为本公开还一实施例的稀疏连接的人工神经网络运算装置的结构示意图;

[0085] 图17是作为本公开还一实施例的映射单元的结构示意图。

具体实施方式

[0086] 为使本公开的目的、技术方案和优点更加清楚明白,以下结合具体实施例,并参照

附图,对本公开作进一步的详细说明。

[0087] 本公开公开了一种用于稀疏人工神经网络的计算装置,包括:

[0088] 映射单元,用于将输入数据转换成输入神经元和权值一一对应的存储方式,并存储在存储装置或者缓存中;

[0089] 存储装置,用于存储数据和指令;

[0090] 运算单元,用于根据所述存储装置中存储的指令对所述数据执行相应的运算;所述运算单元主要执行三步运算,第一步是将输入的神经元和权值数据相乘;第二步执行加法树运算,用于将第一步处理后的加权输出神经元通过加法树逐级相加,或者将加权输出神经元通过和偏置相加得到加偏置输出神经元;第三步执行激活函数运算,得到最终输出神经元。

[0091] 其中,所述映射单元中的一一对应关系表示如下:

[0092] 第一种情形:

[0093] 采用1表示有连接,0表示无连接,每个输出神经元与所有输入神经元的连接状态组成一个0和1的字符串来表示该输出神经元的连接关系;或者

[0094] 采用1表示有连接,0表示无连接,每个输入神经元与所有输出神经元的连接状态组成一个0和1的字符串来表示该输入神经元的连接关系;

[0095] 第二种情形:

[0096] 将一输出神经元第一个连接所在的输入神经元的位置距离第一个输入神经元的距离、所述输出神经元第二个输入神经元距离上一个输入神经元的距离,所述输出神经元第三个输入神经元距离上一个输入神经元的距离,……,依次类推,直到穷举所述输出神经元的所有输入神经元,来表示所述输出神经元的连接关系。

[0097] 作为优选,所述人工神经网络计算装置还包括直接内存存取单元(Direct Memory access,DMA),用于在所述存储装置和缓存中进行数据或者指令读写。

[0098] 作为优选,所述人工神经网络计算装置还包括:

[0099] 指令缓存,用于存储专用指令;以及

[0100] 控制单元,用于从所述指令缓存中读取专用指令,并将其译码成各运算单元指令。

[0101] 作为优选,所述人工神经网络计算装置还包括:

[0102] 输入神经元缓存,用于缓存输入到所述运算单元的输入神经元数据;以及

[0103] 权值缓存,用于缓存权值数据。

[0104] 作为优选,所述人工神经网络计算装置还包括:

[0105] 输出神经元缓存,用于缓存所述运算单元输出的输出神经元。

[0106] 作为优选,所述映射单元用于将输入数据转换成输入神经元和权值一一对应的存储方式,并输出到所述运算单元,而不是存储在存储装置中。

[0107] 作为优选,所述人工神经网络计算装置还包括输入神经元缓存和/或权值缓存,所述输入神经元缓存用于缓存输入到所述运算单元的输入神经元数据,所述权值缓存用于缓存权值数据,所述映射单元用于将输入数据转换成输入神经元和权值一一对应的存储方式,并输出到所述输入神经元缓存和/或权值缓存。

[0108] 作为优选,所述运算单元在第三步执行的激活函数包括sigmoid函数、tanh函数或ReLU函数。

[0109] 本公开还公开了一种用于稀疏连接的人工神经网络的计算方法，包括以下步骤：

[0110] 步骤1，将输入数据转换成输入神经元和权值一一对应的存储方式；其中，所述对应关系包括：

[0111] 第一种情形：

[0112] 采用1表示有连接，0表示无连接，每个输出神经元与所有输入神经元的连接状态组成一个0和1的字符串来表示该输出神经元的连接关系；或者

[0113] 采用1表示有连接，0表示无连接，每个输入神经元与所有输出神经元的连接状态组成一个0和1的字符串来表示该输入神经元的连接关系；

[0114] 第二种情形：

[0115] 将一输出神经元第一个连接所在的输入神经元的位置距离第一个输入神经元的距离、所述输出神经元第二个输入神经元距离上一个输入神经元的距离，所述输出神经元第三个输入神经元距离上一个输入神经元的距离，……，依次类推，直到穷举所述输出神经元的所有输入神经元，来表示所述输出神经元的连接关系

[0116] 步骤2，将输入的神经元和权值数据相乘；

[0117] 步骤3，执行加法树运算，将第一步处理后的加权输出神经元通过加法树逐级相加，或者将加权输出神经元通过和偏置相加得到加偏置输出神经元；

[0118] 步骤4，对加权输出神经元或加偏置输出神经元执行激活函数运算，得到最终输出神经元；其中，所述激活函数包括sigmoid函数、tanh函数或ReLU函数。

[0119] 下面结合附图和具体实施例对本公开的技术方案进一步阐释说明。

[0120] 图3是根据本公开一个实施例的总体结构的示意性框图。

[0121] I/O接口1，用于I/O数据需要经过CPU3发给稀疏的多层人工神经网络运算装置，然后由稀疏的多层人工神经网络运算装置4写入存储装置，稀疏的多层人工神经网络运算装置4需要的专用程序也是由CPU3传输到稀疏的多层人工神经网络运算装置4。

[0122] 存储装置2用于暂存稀疏的多层人工神经网络模型和神经元数据，特别是当全部模型无法在稀疏的多层人工神经网络运算装置4上的缓存中放下时。

[0123] 中央处理器CPU3，用于进行数据搬运以及稀疏的多层人工神经网络运算装置4启动停止等基本控制，作为稀疏的多层人工神经网络运算装置4与外部控制的接口。

[0124] 稀疏的人工神经网络运算装置4，用于执行稀疏的多层人工神经网络运算单元，接受来自CPU3的数据和程序，执行上述稀疏的多层人工神经网络运算算法，稀疏的人工神经网络运算装置4的执行结果将传输回CPU3。

[0125] 通用系统结构：将稀疏的人工神经网络运算装置4作为CPU 3或者GPU的协处理器来执行稀疏的多层人工神经网络运算算法。

[0126] 多个稀疏的人工神经网络运算装置互联的系统结构：多个稀疏的人工神经网络运算装置4可以通过PCIE总线互联，以支持更大规模的稀疏的多层人工神经网络运算，可以共用同一个宿主CPU或者分别有自己的宿主CPU，可以共享内存也可以每个加速器有各自的内存。此外其互联方式可以是任意互联拓扑。

[0127] 对于一个稀疏连接的神经网络如图4所示，有4个输入神经元：i_1，i_2，i_3，i_4，有2个输出神经元：o_1，o_2。其中，o_1和i_1，i_3，i_4有连接，把连接的权值分别表示为w_{11}，w_{31}，w_{41}，o_2和i_2，i_3有连接，把连接的权值分别表示为w_{22}，w_{32}。

[0128] 有两种方法可以表示上面稀疏神经网络的连接关系，一种是每个输入与输出神经元之间都用一位表示是否有连接，另一种是用连接之间的距离来表示每个连接的位置。

[0129] 第一种连接表示：

[0130] 对于图4中的神经网络，如图5所示，输出神经元o_1的连接关系为：1011，每一位表示是否与输入神经元有连接，1表示有连接，0表示无连接，输出神经元o_2的连接关系为0110。在运算时，连接关系为0所对应的输入神经元不会进行运算。

[0131] 在存储连接关系时，可以按照优先输入神经元或者输出神经元的顺序对连接关系进行存储。具体存储格式有以下几种：

[0132] 格式一：将每个输出神经元的所有输入神经元依次摆放完，上面的例子摆放的顺序为10110110。

[0133] 格式二：将每个输入神经元的所有的输出神经元依次摆放完，上面的例子摆放的顺序为10011110。

[0134] 第二种连接表示：

[0135] 比如对于图6中的神经网络，输出神经元o_1与输入神经元i_1，i_3，i_4相连接，那么连接关系为0，2，1。0表示第一个连接所在的位置距离第一个输入神经元的距离为0，即第一个输入神经元，2表示第二个输入神经元距离上一个输入神经元的距离为2，即表示第三个输入神经元，1表示第三个输入神经元距离上一个输入神经元的距离为1，即表示第四个输入神经元。同理，o_2的连接关系为1，1。

[0136] 本公开的映射单元包括但不限于以上的连接关系。

[0137] 卷积神经网络是人工神经网络的一种，卷积层包含多个滤波器，也就是卷积核，这些卷积核重复的作用于所有输入图像上，提取局部特征。不同的卷积核能够提取出不同种类的局部特征，一副输入图像在经过卷积层之后就变成一些能够被更好理解的抽象特征。

[0138] 自然图像有其固有特性，也就是说，图像的一部分的统计特性与其他部分是一样的。这也意味着在这一部分学习的特征也能用在另一部分上，所以对于这个图像上的所有位置，都能使用同样的学习特征。当从一个大尺寸图像中随机选取一小块，比如说8*8作为样本，并且从这个小块样本中学习到了一些特征，这时可以把从这个8*8样本中学习到的特征作为探测器，应用到这个图像的任意地方中去。特别是，可以用从8*8样本中学习到的特征跟原本的大尺寸图像做卷积，从而对这个大尺寸图像上的任意位置获得一个不同特征的激活值。这个8*8的样本特征被称作卷积核。

[0139] 如图7是一个卷积操作的例子。卷积核是一个2*2的矩阵，卷积核在输入图像上滑动。

[0140] 假设每次滑动一个像素点，则总共会有四次卷积操作。对于每次卷积操作，卷积核矩阵与对应的输入图像数据做乘加操作。

[0141] 假设卷积核的权值变得稀疏，由之前的2*2，变成只有两个参数，如图8所示。则对于输出o_0来说，需要的输入神经元为i_0，i_1，i_3，i_4，输入权值为：w_0，w_3，连接关系为1001或者0，2；

[0142] 对于输出o_3来说，需要的输入神经元为i_3，i_5，i_7，i_8，输入权值为：w_0，w_3，连接关系为1001或者0，2。

[0143] 由此可见，对于同个输出特征图上的不同的输出神经元，所需要的输入神经元不

同,权值和连接关系是相同的。

[0144] 可执行稀疏连接的人工神经网络运算装置可以处理各种稀疏连接表示的稀疏连接的人工神经网络,可执行稀疏连接的人工神经网络运算装置中有一个专门用于处理稀疏连接的单元,在这里称为映射单元,对于不同的稀疏连接关系和处理方法,稀疏连接的人工神经网络运算装置结构会略有不同,下面将分别描述不同的结构和方法。

[0145] 结构和方法一

[0146] 如图9所示,映射单元5,用来将输入数据转换成输入神经元和权值一一对应的存储方式。

[0147] 存储装置1,用来存储数据和指令,尤其是神经网络规模很大的时候,指令缓存3、输入神经元缓存6、输出神经元缓存9、权值缓存8放不下这么多数据,只能将数据临时存放在存储装置1。

[0148] 直接内存存取单元DMA2,用来将存储装置中的数据或者指令搬到各个缓存中。

[0149] 指令缓存3,用来存储专用指令。

[0150] 控制单元4,从指令缓存3中读取专用指令,并将其译码成各运算单元指令。

[0151] 输入神经元缓存6,用来存储运算的输入神经元数据。

[0152] 运算单元7,用于执行具体的运算。运算单元包括三个阶段,第一阶段执行乘法运算,用于将输入的神经元和权值数据相乘。第二阶段执行加法树运算,第一、二两阶段合起来完成了向量内积运算。第三阶段执行激活函数运算,激活函数可以是sigmoid函数、tanh函数等。第三阶段得到输出神经元,写回到输出神经元缓存。

[0153] 权值缓存8,用来存储权值数据。

[0154] 输出神经元缓存9,用来存储运算的输出神经元。

[0155] 映射单元的结构如图10所示。

[0156] 以上面稀疏连接的神经网络为例,连接关系可以是上述的两种稀疏表示之一,映射单元会根据连接关系,将输入神经元和输入权值按照连接关系输出映射后的神经元和权值,映射后的神经元和权值可以在运算时被直接使用而不需要考虑连接关系,对于输出神经元o1映射的具体过程如下:输入神经元为:i_1,i_2,i_3,i_4,输入权值为:w_{11},w_{31},w_{41},连接关系可以为:1011,或0,2,1。映射单元根据连接关系,将输入神经元和权值变成相对应的关系,输出包括两种情况:一种是去除掉没有连接的输入神经元,则映射后的神经元为i_1,i_3,i_4,映射后的权值为w_{11},w_{31},w_{41};另一种是权值在没有连接的地方补成0,则映射后的神经元为i_1,i_2,i_3,i_4,映射后的权值为w_{11},0,w_{31},w_{41}。

[0157] 运算单元包括三个部分,第一部分乘法器,第二部分加法树,第三部分为非线性函数单元。第一部分将输入神经元(in)通过和权值(w)相乘得到加权输出神经元(out),过程为:out=w*in;第二部分将加权输出神经元通过加法树逐级相加,另外还可以将输出神经元(in)通过和偏置(b)相加得到加偏置输出神经元(out),过程为:out=in+b;第三部分将输出神经元(in)通过激活函数(active)运算得到激活输出神经元(out),过程为:out=active(in),激活函数active可以是sigmoid、tanh、relu、softmax等,除了做激活操作,第三部分可以实现其他的非线性函数,可将输入神经元(in)通过运算(f)得到输出神经元(out),过程为:out=f(in)。

[0158] 运算过程如图11所示。

[0159] 结构和方法二

[0160] 如图12所示，存储装置1，用来存储数据和指令，尤其是神经网络规模很大的时候，指令缓存3、输入神经元缓存6、输出神经元缓存9、权值缓存8放不下这么多数据，只能将数据临时存放在存储装置1。

[0161] 直接内存存取单元DMA2，用来将存储装置中的数据或者指令搬到各个缓存中。

[0162] 指令缓存3，用来存储专用指令。

[0163] 控制单元4，从指令缓存3中读取专用指令，并将其译码成各运算单元指令。

[0164] 映射单元5，用来将输入数据转换成输入神经元和权值一一对应的存储方式。

[0165] 输入神经元缓存6，用来存储运算的输入神经元数据。

[0166] 运算单元7，用于执行具体的运算。运算单元包括三个阶段，第一阶段执行乘法运算，用于将输入的神经元和权值数据相乘。第二阶段执行加法树运算，第一、二两阶段合起来完成了向量内积运算。第三阶段执行激活函数运算，激活函数可以是sigmoid函数、tanh函数等。第三阶段得到输出神经元，写回到输出神经元缓存。

[0167] 权值缓存8，用来存储权值数据。

[0168] 输出神经元缓存9，用来存储运算的输出神经元。

[0169] 映射单元的结构如图13所示。

[0170] 以上述稀疏连接的神经网络为例，连接关系可以是上述的两种稀疏表示之一，映射单元会根据连接关系，将输入神经元和输入权值按照连接关系输出映射后的神经元和权值，映射后的神经元和权值可以在运算时被直接使用而不需要考虑连接关系，对于输出神经元o1映射的具体过程如下：

[0171] 输入神经元为：i_1,i_2,i_3,i_4，输入权值为：w_{11},w_{31},w_{41}，连接关系可以为：1011，或0，2，1。映射单元根据连接关系，将输入神经元和权值变成相对应的关系，输出有两种情况：一种是去除掉没有连接的输入神经元，则映射后的神经元为i_1,i_3,i_4，映射后的权值为w_{11},w_{31},w_{41}；另一种是权值在没有连接的地方补成0，则映射后的神经元为i_1,i_2,i_3,i_4，映射后的权值为$w_{11},0,w_{31},w_{41}$。

[0172] 结构和方法一和结构方法二中的映射单元的主要区别是结构和方法一中的映射单元是在计算之前事先把输入神经元和权值映射好后存储在存储装置中，结构和方法二是在计算中进行映射，将映射好的数据直接给运算单元进行运算。

[0173] 结构和方法三：

[0174] 基于结构和方法二稍作修改可以改成如图14所示的结构，映射单元只对输入神经元进行映射。

[0175] 此时，映射单元的结构图如图15所示。

[0176] 对于输出神经元o_1映射的具体过程如下：

[0177] 输入神经元为：i_1,i_2,i_3,i_4，连接关系可以为：1011，或者：0，2，1。映射单元根据连接关系，将输入神经元和权值变成相对应的关系，去除掉没有连接的输入神经元，则映射后的神经元为i_1,i_3,i_4。

[0178] 结构和方法四：

[0179] 基于结构和方法二稍作修改可以改成如图16所示的结构，映射单元只对输入权值进行映射。

[0180] 此时，映射单元的结构图如图17所示。

[0181] 对于输出神经元o_1映射的具体过程如下：

[0182] 输入权值为：w_{11}，w_{31}，w_{41}，连接关系可以为：1011，或者：0，2，1。映射单元根据连接关系，将输入神经元和权值变成相对应的关系，映射后的权值为w_{11}，0，w_{31}，w_{41}。

[0183] 以上所述的具体实施例，对本公开的目的、技术方案和有益效果进行了进一步详细说明，应理解的是，以上所述仅为本公开的具体实施例而已，并不用于限制本公开，凡在本公开的精神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本公开的保护范围之内。

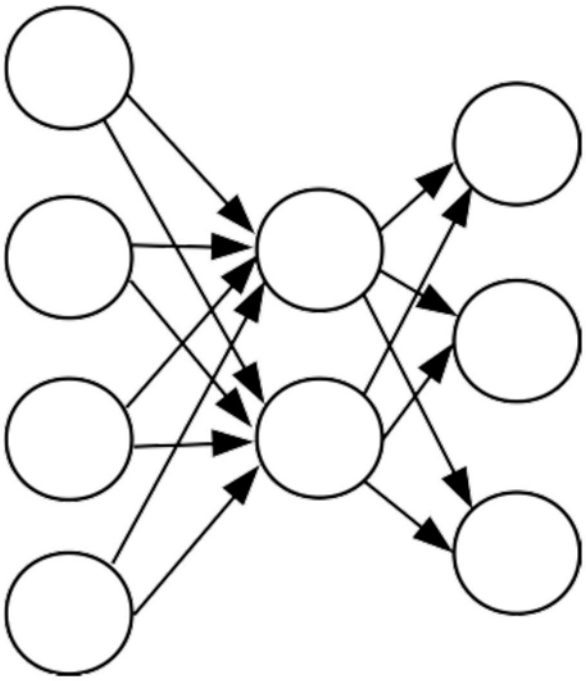


图1

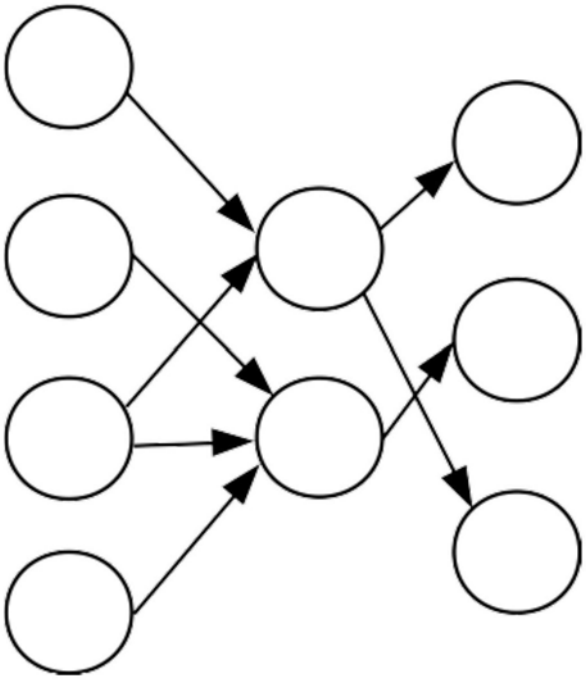


图2

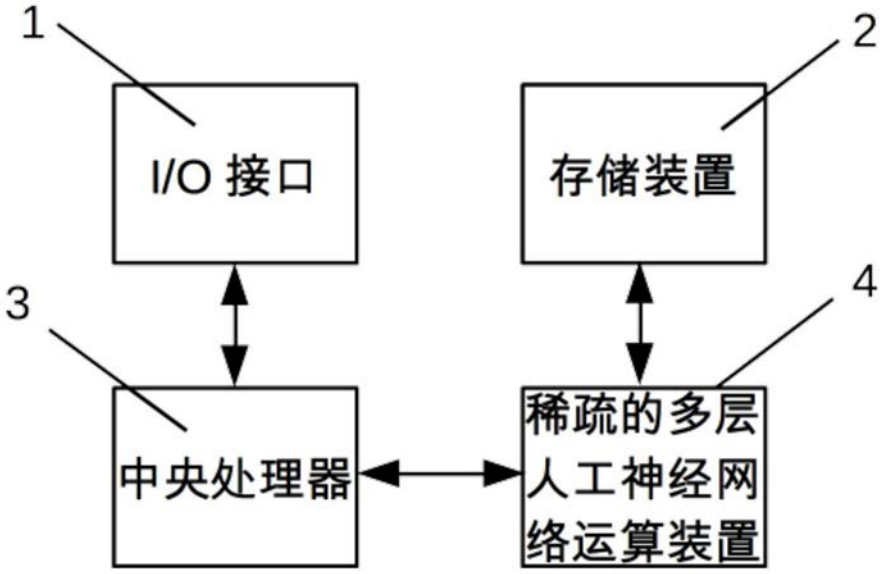


图3

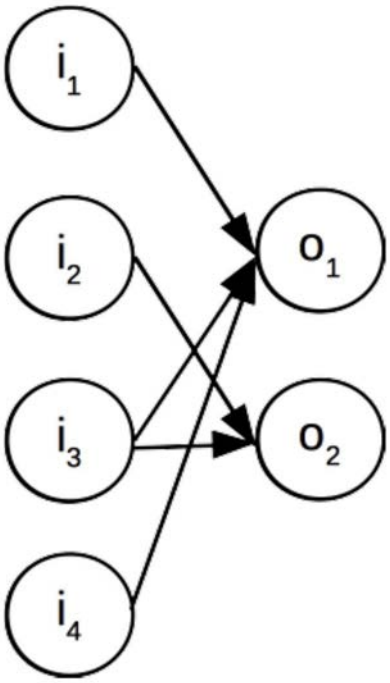


图4

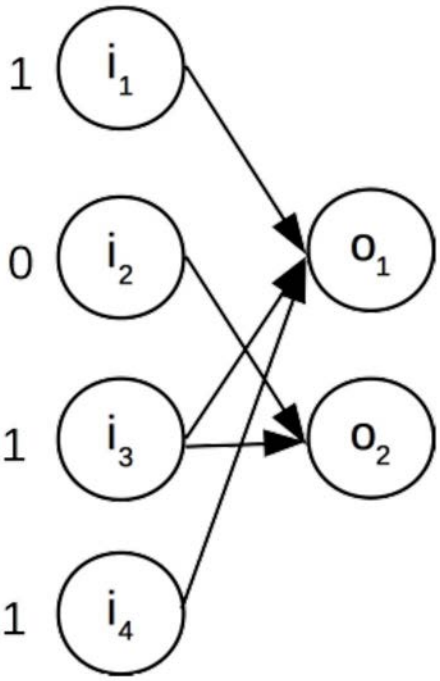


图5

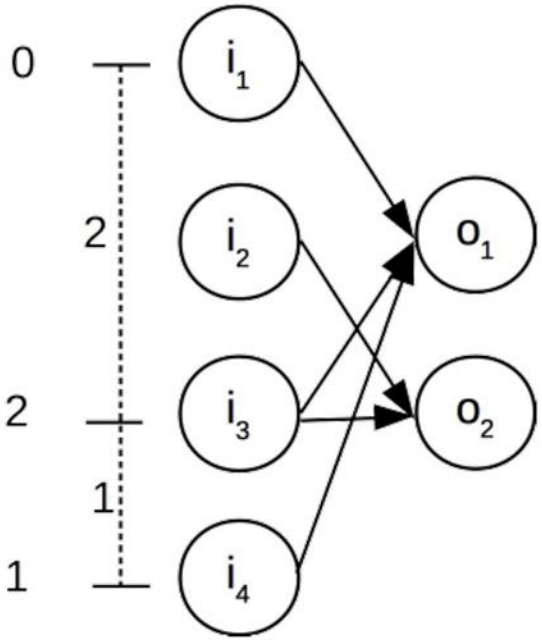


图6

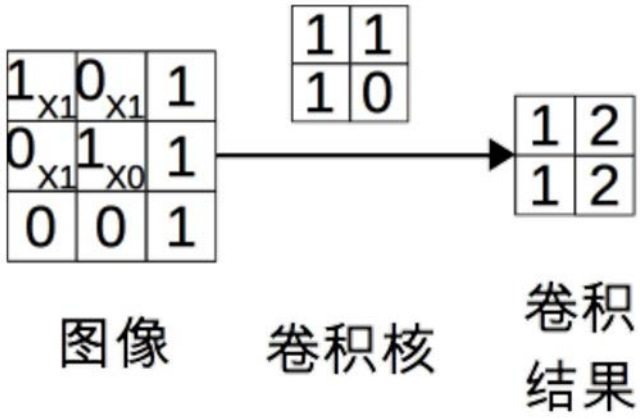


图7

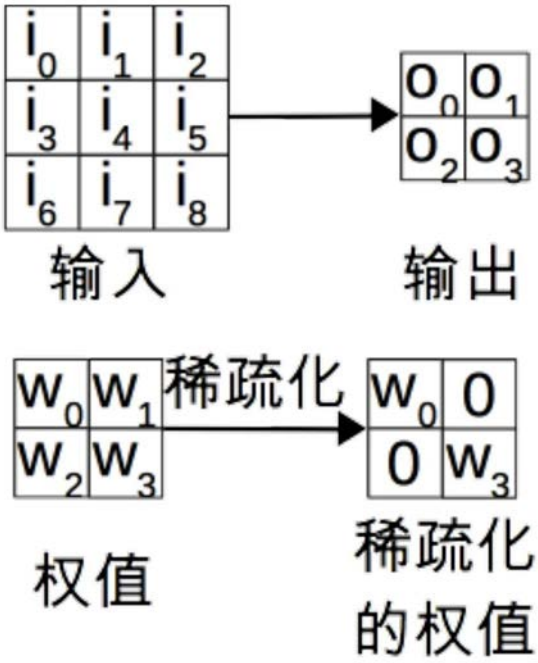


图8

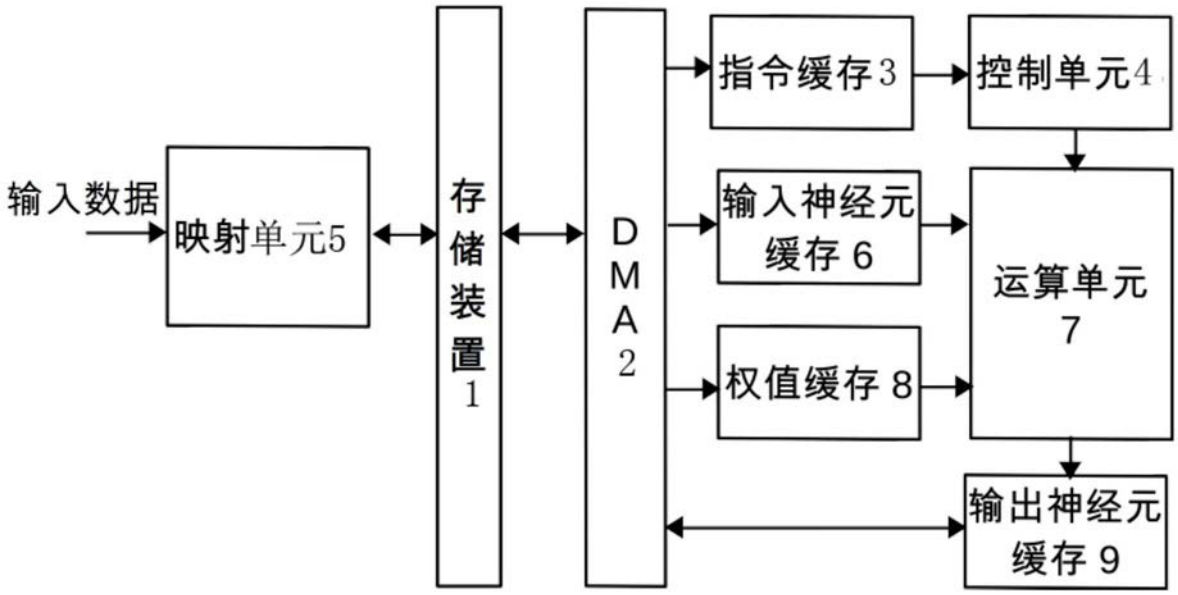

输入数据
映射单元5
存储装置1
DMA2
指令缓存3
控制单元4
输入神经元缓存6
运算单元7
权值缓存8
输出神经元缓存9

图9

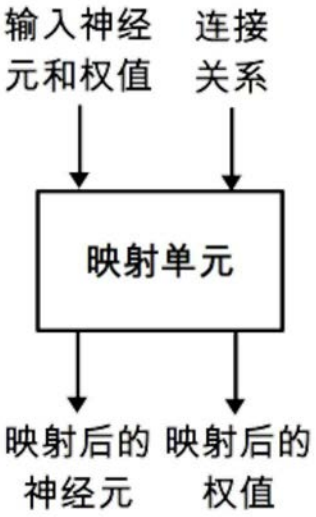

输入神经元和权值
连接关系
映射单元
映射后的神经元
映射后的权值

图10

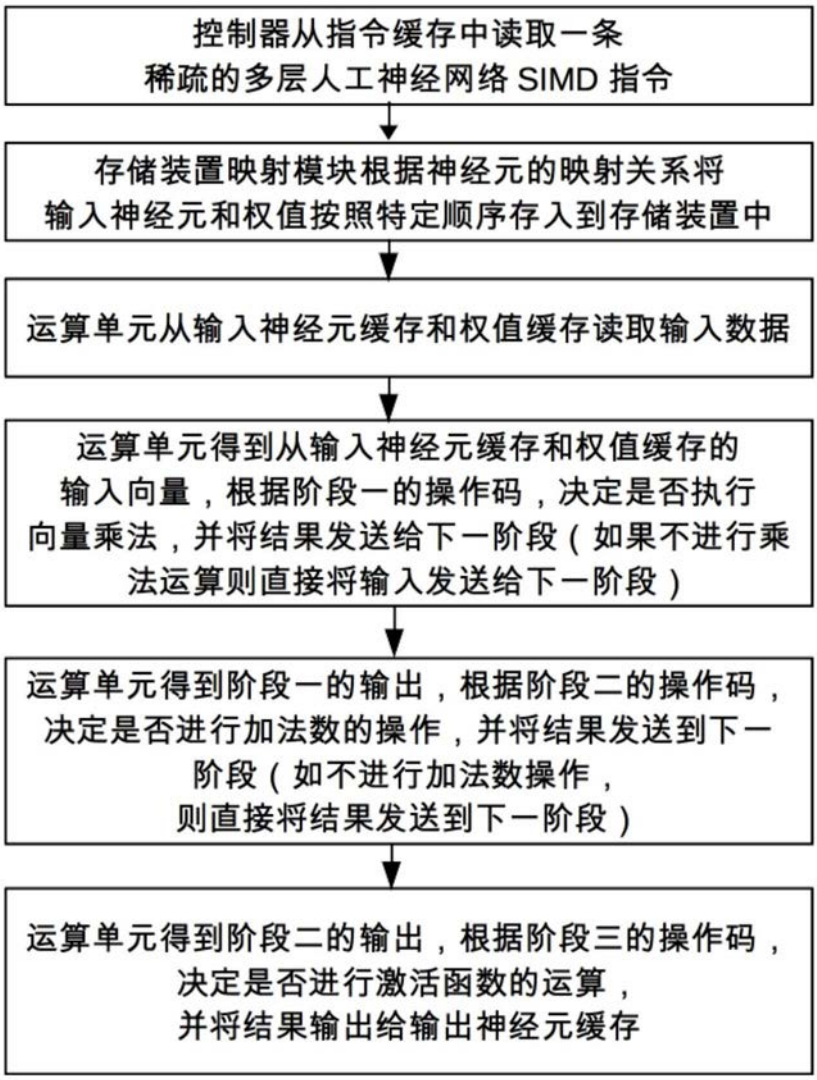

控制器从指令缓存中读取一条
稀疏的多层人工神经网络 SIMD 指令
存储装置映射模块根据神经元的映射关系将
输入神经元和权值按照特定顺序存入到存储装置中
运算单元从输入神经元缓存和权值缓存读取输入数据
运算单元得到从输入神经元缓存和权值缓存的
输入向量，根据阶段一的操作码，决定是否执行
向量乘法，并将结果发送给下一阶段（如果不进行乘
法运算则直接将输入发送给下一阶段）
运算单元得到阶段一的输出，根据阶段二的操作码，
决定是否进行加法数的操作，并将结果发送到下一
阶段（如不进行加法数操作，
则直接将结果发送到下一阶段）
运算单元得到阶段二的输出，根据阶段三的操作码，
决定是否进行激活函数的运算，
并将结果输出给输出神经元缓存

图11

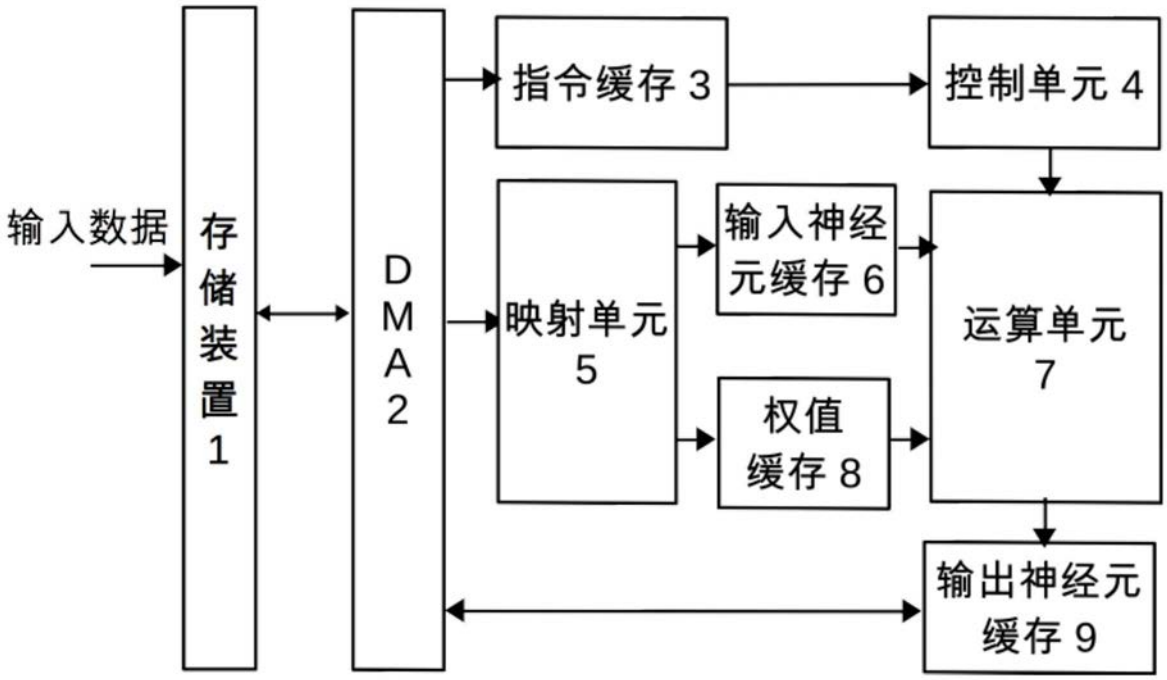

输入数据
存储装置1
DMA2
指令缓存3
控制单元4
映射单元5
输入神经元缓存6
权值缓存8
运算单元7
输出神经元缓存9

图12

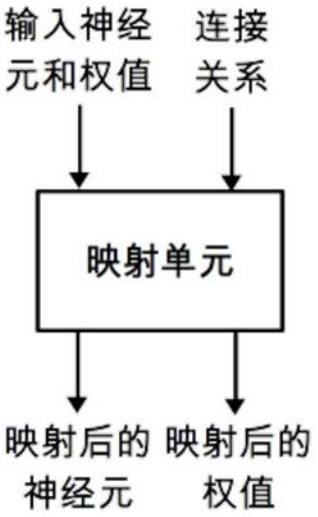

输入神经元和权值
连接关系
映射单元
映射后的神经元
映射后的权值

图13

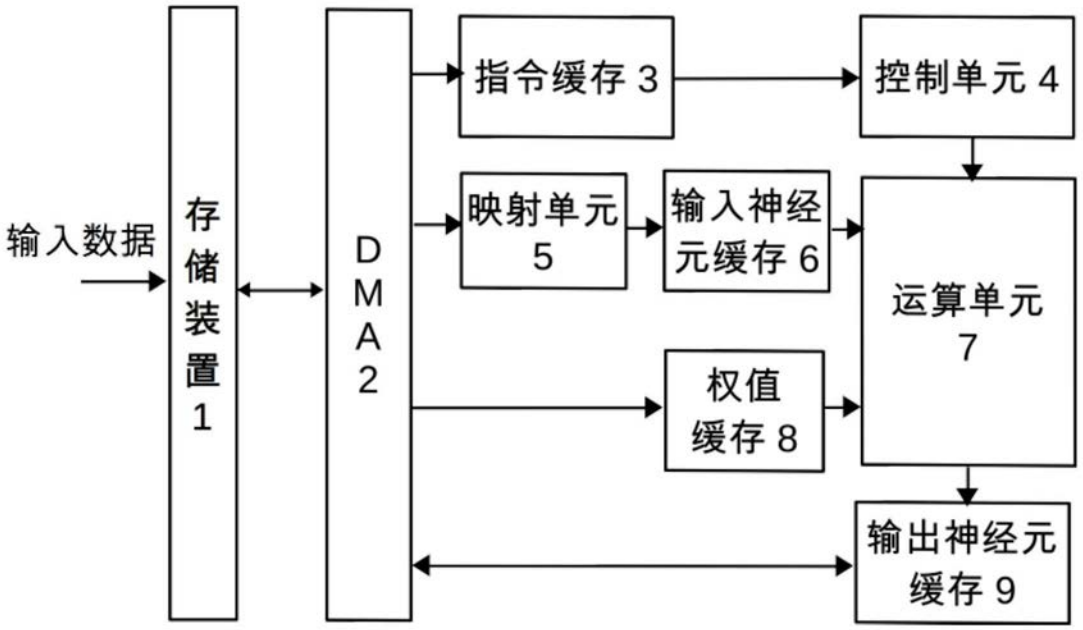

输入数据
存储装置1
DMA2
指令缓存3
控制单元4
映射单元5
输入神经元缓存6
运算单元7
权值缓存8
输出神经元缓存9

图14

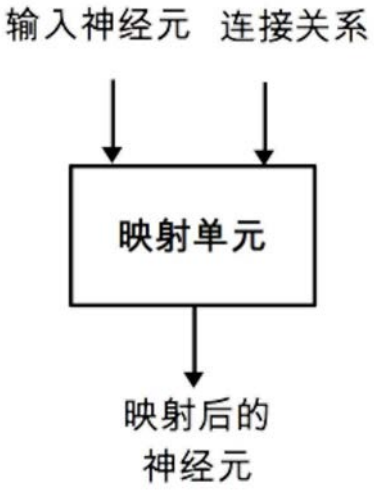

输入神经元 连接关系
映射单元
映射后的神经元

图15

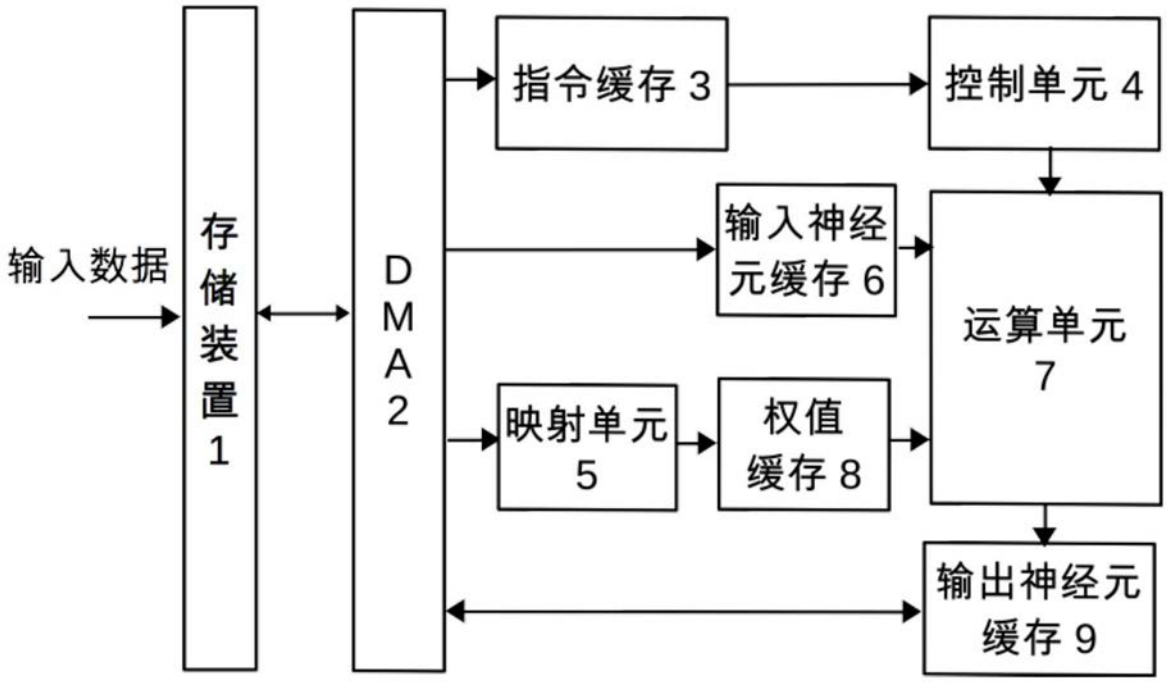

输入数据
存储装置1
DMA2
指令缓存3
控制单元4
输入神经元缓存6
运算单元7
映射单元5
权值缓存8
输出神经元缓存9

图16

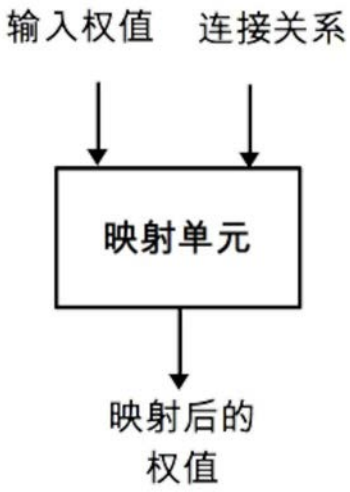

输入权值
连接关系
映射单元
映射后的权值

图17