(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2017 年 10 月 19 日 (19.10.2017)

WIPO | PCT

(10) 国际公布号
WO 2017/177442 A1

(51) 国际专利分类号:
G06N 3/02 (2006.01)

(21) 国际申请号: PCT/CN2016/079431

(22) 国际申请日: 2016 年 4 月 15 日 (15.04.2016)

(25) 申请语言: 中文

(26) 公布语言: 中文

(71) 申请人: **北京中科寒武纪科技有限公司 (CAMBRICON TECHNOLOGIES CO., LTD.)** [CN/CN]; 中国北京市海淀区科学院南路 6 号科研综合楼 644 室, Beijing 100190 (CN)。

(72) 发明人: **刘少礼 (LIU, Shaoli)**; 中国北京市海淀区科学院南路 6 号科研综合楼 644 室, Beijing 100190 (CN)。 **于涌 (YU, Yong)**; 中国北京市海淀区科学院南路 6 号科研综合楼 644 室, Beijing 100190 (CN)。 **陈云霁 (CHEN, Yunji)**; 中国北京市海淀区科学院南路 6 号科研综合楼 644 室, Beijing 100190 (CN)。 **陈天石 (CHEN, Tianshi)**; 中国北京市海淀区科学院南路 6 号科研综合楼 644 室, Beijing 100190 (CN)。

(74) 代理人: **中科专利商标代理有限责任公司 (CHINA SCIENCE PATENT & TRADEMARK AGENT LTD.)**; 中国北京市海淀区西三环北路 87 号 4-1105 室, Beijing 100089 (CN)。

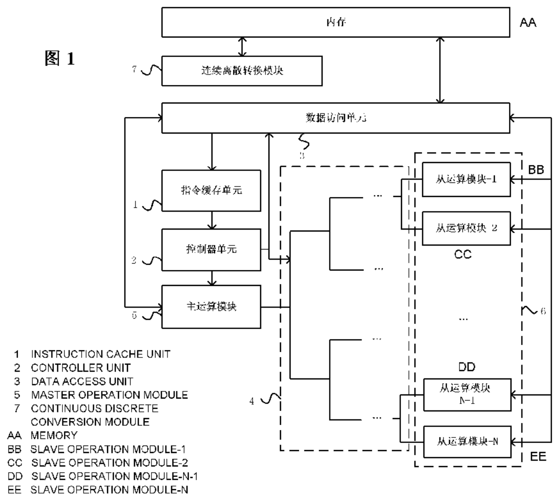
(81) 指定国 (除另有指明，要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA,

[见续页]

(54) Title: DISCRETE DATA REPRESENTATION SUPPORTED DEVICE AND METHOD FOR FORWARD OPERATION OF ARTIFICIAL NEURAL NETWORK

(54) 发明名称 : 支持离散数据表示的人工神经网络正向运算装置和方法

图 1



(57) Abstract: The present invention provides a discrete data representation supported device for executing forward operation of an artificial neural network, comprising an instruction cache unit, a controller unit, a data access unit, an interconnection module, a master operation module, a plurality of slave operation modules, a discrete data operation module, and a continuous discrete conversion module. By means of the device, the forward operation of a multi-layer artificial neural network supporting discrete data representation can be implemented. Data, such as a weight value and a neuron, in the forward operation process can be represented in a discrete form. For example, -1, -1/2, 0, 1/2, and 1 are discontinuous data. A module for supporting discrete data operation is provided. Different bitwise operations, such as data Exclusive-OR and NOT, are used according to values of discrete data to replace basic operations, such as multiplication and addition, of continuous data. A module for converting continuous data into discrete data is provided. The device is provided for supporting batch normalization operation.

(57) 摘要:

[见续页]

WO 2017/177442 A1

RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国 (除另有指明，要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第 21 条(3))。

本发明提供了一种支持离散数据表示的用于执行人工神经网络正向运算的装置，包括指令缓存单元、控制器单元、数据访问单元、互联模块、主运算模块、以及多个从运算模块、离散数据运算模块、连续离散转换模块。使用该装置可以实现支持离散数据表示的多层人工神经网络的正向运算。正向运算过程中的权值、神经元等数据可以采用离散形式表示。例如-1、-1/2、0、1/2、1 等不是连续的数据。提供了支持离散数据运算的模块。根据离散数据的值采用不同位运算例如数据的异或、取非等代替连续数据的基本运算例如乘法、加法等。提供了将连续数据转换为离散数据的模块。提供了利用上述装置支持批归一化(batch normalization)计算。

支持离散数据表示的人工神经网络正向运算装置和方法

技术领域

本发明总体上涉及人工神经网络，具体地涉及一种用于执行人工神经网络正向运算的装置和方法，本发明中的数据支持离散数据表示。并对离散数据，用按位运算例如异或、取非等代替了连续数据基本运算例如乘法等操作。

背景技术

多层人工神经网络被广泛应用于模式识别，图像处理，函数逼近和优化计算等领域，多层人工网络在近年来由于其较高的识别准确度和较好的可并行性，受到学术界和工业界越来越广泛的关注。

一种支持多层人工神经网络正向运算的已知方法是使用通用处理器。该方法通过使用通用寄存器堆和通用功能部件执行通用指令来支持上述算法。另一种支持多层人工神经网络正向训练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。

这两种装置在数据存储和运算上都是使用的连续数据。连续数据的存储需要较多的资源，例如一个 32 位的浮点数据，就需要 32 个比特位来存储该数据。在连续数据的运算上，所需要的加法器、乘法器等功能部件的实现也较为复杂。

离散数据表示指通过特定的数字来代替连续数据的存储方式。例如，可以通过 00、01、10、11 四个数字分别代表数据-1、-1/8、1/8、1 四个数字。这种存储方式不同于连续存储方式。在连续存储方式中二进制数 00/01/10/11 就代表着十进制中 0/1/2/3 四个连续的数字。通过这种类似于索引的表示方式我们用形式上连续的数字代替了不连续离散化的真实数据。存储的数字不是连续的，所以叫离散数据表示。

目前的运算多层人工神经网络的运算装置在数据表示上已知方法是用浮点数或者定点数这样的连续数据表示。因为多层神经网络权值的精度较高和数量较大，连续数据的表示方式在运算和存储两方面带来更大的开销。而通过离散数据表示的方式，可以通过数据按位的异或、移位等运算代替连续数据的乘法等运算。从而大大减少乘法器部件的数量。并且用几个比特位的离散化数据，相比于传统

的 32 位的浮点数存储上的优势也是明显的。

发明内容

本发明的一个方面提供了一种支持离散数据表示的用于执行人工神经网络正向运算的装置，包括指令缓存单元、控制器单元、数据访问单元、互联模块、主运算模块、多个从运算模块，其中：

指令缓存单元用于通过数据访问单元读入指令并缓存读入的指令；

控制器单元用于从指令缓存单元读取指令，并将该指令译码成控制互联模块、主运算模块、以及从运算模块行为的微指令；

数据访问单元用于从外部地址空间向主运算模块和各从运算模块的相应数据缓存单元中写入离散数据或连续数据或从所述数据缓存单元向外部地址空间读离散数据或连续数据；

每层神经网络开始正向计算的阶段，主运算模块通过互联模块向所有的从运算模块传输本层的离散或连续的输入神经元向量，在从运算模块的计算过程完成后，互联模块逐级将各从运算模块的离散或连续化的输出神经元值拼成中间结果向量，其中，当输入数据是离散数据与连续数据的混合数据时，从运算模块针对不同离散数据采取预先设置的相应计算方式；

主运算模块用于利用中间结果向量完成后续计算，当输入数据是离散数据与连续数据的混合数据时，主运算模块针对不同离散数据采取预先设置的相应计算方式。

可选地，离散数据表示指用特定的离散数字代替真实的连续数据的表示方式。

可选地，多个从运算模块利用相同的离散或连续的输入神经元向量和各自不同的离散或连续的权值向量，并行地计算出各自的离散或连续输出神经元值。

可选地，主运算模块对中间结果向量执行以下任一项操作：

加偏置操作，在中间结果向量上加上偏置；

对中间结果向量进行激活，激活函数 active 是非线性函数 sigmoid，tanh，relu，softmax 中的任一个或线性函数；

采样操作，将中间结果向量与随机数比较，大于随机数则输出 1，小于随机

数则输出 0；或者

池化操作，包括最大值池化或平均值池化。

可选地，从运算模块包括输入神经元缓存单元，用于缓存离散或者连续的输入神经元向量。

可选地，互联模块构成主运算模块和所述多个从运算模块之间的连续或离散化数据的数据通路，可以实现成不同的互联拓扑。在一种实施方式中，具有 H 树型的结构，H 树是由多个节点构成的二叉树通路，每个节点将上游的离散或连续的数据同样地发给下游的两个节点，将下游的两个节点返回的离散或连续的数据合并，并返回给上游的节点。

可选地，主运算模块包括运算单元、数据依赖关系判断单元和神经元缓存单元，其中：

神经元缓存单元用于缓存主运算模块在计算过程中用到的离散或连续表示的输入数据和输出数据；

运算单元完成主运算模块的各种运算功能，当输入数据是离散数据与连续数据的混合数据时，针对不同离散数据采取预先设置的相应计算方式；

数据依赖关系判断单元是运算单元读写神经元缓存单元的端口，保证对神经元缓存单元中连续数据或离散数据读写不存在一致性冲突，并且负责从神经元缓存单元读取输入离散或连续的神经元向量，并通过互联模块发送给从运算模块；以及

来自互联模块的中间结果向量被发送到运算单元。

可选地，每个从运算模块包括运算单元、数据依赖关系判定单元、神经元缓存单元和权值缓存单元，其中：

运算单元接收控制器单元发出的微指令并进行算数逻辑运算，当输入数据是离散数据与连续数据的混合数据时，针对不同离散数据采取预先设置的相应计算方式；

数据依赖关系判断单元负责计算过程中对支持离散数据表示的神经元缓存单元和支持离散数据表示的权值缓存单元的读写操作，保证对支持离散数据表示的神经元缓存单元和支持离散数据表示的权值缓存单元的读写不存在一致性冲突；

神经元缓存单元缓存输入神经元向量的数据以及该从运算模块计算得到的输出神经元值；以及

权值缓存单元缓存该从运算模块在计算过程中需要的离散或连续表示的权值向量。

可选地，数据依赖关系判断单元通过以下方式保证读写不存在一致性冲突：判断尚未执行的微指令与正在执行过程中的微指令的数据之间是否存在依赖关系，如果不存在，允许该条微指令立即发射，否则需要等到该条微指令所依赖的所有微指令全部执行完成后该条微指令才允许被发射。

可选地，主运算模块或从运算模块中的运算单元包括运算决定单元和混合数据运算单元，当输入数据是混合数据时，运算决定单元根据其中的离散数据决定应对该混合数据执行何种操作，然后，混合数据运算单元根据运算决定单元的决定结果，执行相应操作。

可选地，主运算模块或从运算模块中的所述运算单元还包括离散数据运算单元和连续数据运算单元中的至少一个，以及数据类型判断单元，当输入数据全是离散数据时，由离散数据运算单元根据输入的离散数据通过查表执行相应操作，当输入数据全是连续数据时，由连续数据运算单元执行相应操作。

可选地，该装置还包括连续离散转换单元，连续离散转换单元包括预处理模块、距离计算模块、和判断模块，假设使用 M 个离散数据， $M=2m$，$m \geq 1$，令这些离散数据分别对应于预定区间[-zone,zone]内的 M 个数值，其中：

预处理模块对于输入的连续数据 x 使用 clip(-zone,zone)运算进行预处理，得到区间[-zone,zone]内的预处理数据 y，其中，如果 $x \leq -zone$ 则 $y = -zone$，如果 $x \geq zone$ 则 $y = zone$，如果 $-zone < x < zone$，则预处理数据 $y = x$；

距离计算模块计算预处理数据 y 与上述各数值之间的距离；以及

判断模块基于该距离计算并输出离散数据。

可选地，预定区间[-zone,zone]是[-1,1]或[-2,2]；并且/或者 M 个数值的绝对值是 2 的幂的倒数；并且/或者判断模块执行：输出与该预处理数据 y 距离最近的数值所对应的离散数据，如果有两个数值与该预处理数据距离相等，则输出二者中任一个所对应的离散数据；或者计算预处理数据 y 分别到距离最近的两个数值中任一个的归一化概率，将这两个数值中任一个所对应的归一化概率与随机数生成

模块生成的(0,1)之间的随机数 z 比较，如果该 z 小于该概率则输出该离散数据，否则输出另一离散数据。

本发明的另一个方面提供了一种使用上述装置执行单层人工神经网络正向运算的方法。通过提供的指令集，控制器控制读入运算所需的神经元、权值、常数量等数据。这些数据可以采用离散数据表示也可以不采用。之后主运算模块、从运算模块、互联模块完成神经元数据与权值数据相乘加偏置激活的过程。特别对于离散数据表示的数据，在进行乘法操作时，根据离散数据的数值，通过对相关数据的位操作代替了乘法运算。例如权值数据用 1 比特的离散数据表示，0 代表+1,1 代表-1，通过对与权值相乘数据的符号位异或，实现了对权值的乘法运算。

本发明的另一个方面提供了一种使用上述装置支持人工神经网络批归一化运算(Batch Normalization)的方法。通过提供的指令集，控制器控制数据访问单元读入输入的数据，之后控制主从运算模块根据 batch 大小求出各自位置的均值以及方差或使用设定好的均值方差。之后控制器控制对应位置的输入数据减去均值除以方差。最后控制器控制用处理后的数据与学习参数相乘后加上另一个学习参数。

本发明的另一方面提供了一种使用上述装置执行多层人工神经网络正向运算的方法。其实现过程与单层神经网络类似，当上一层人工神经网络执行完毕后，下一层的运算指令会将主运算单元中存储的上一层的输出神经元地址作为本层的输入神经元地址。同样地，指令中的权值地址和偏置地址也会变更至本层对应的地址。

附图说明

为了更完整地理解本发明及其优势，现在将参考结合附图的以下描述，其中：

图 1 示出了根据本发明实施例的用于执行支持离散数据表示的人工神经网络正向运算的装置的整体结构的示例框图。

图 2 示意性示出了根据本发明实施例的用于执行支持离散数据表示的人工神经网络正向运算的装置中 H 树模块（互联模块的一种实施方式）的结构。

图 3 示出了根据本发明实施例的用于执行支持离散数据表示的人工神经网络正向运算的装置中主运算模块结构的示例框图。

图4示出了根据本发明实施例的用于执行支持离散数据表示的人工神经网络正向运算的装置中从运算模块结构的示例框图。

图5示出了根据本发明实施例的神经网络正向运算过程的示例框图。

图6示出了根据本发明实施例的支持离散数据表示的神经网络反向训练过程的示例框图。

图7示出了根据本发明实施例的单层人工神经网络运算的流程图。

图8示出了根据本发明实施例的运算单元示例结构。

图9示出了根据本发明实施例的连续数据和离散数据转化的连续离散转化模块的示例结构。

在所有附图中，相同的装置、部件、单元等使用相同的附图标记来表示。

具体实施方式

根据结合附图对本发明示例性实施例的以下详细描述，本发明的其它方面、优势和突出特征对于本领域技术人员将变得显而易见。

在本发明中，术语“包括”和“含有”及其派生词意为包括而非限制；术语“或”是包含性的，意为和/或。

在本说明书中，下述用于描述本发明原理的各种实施例只是说明，不应该以任何方式解释为限制发明的范围。参照附图的下述描述用于帮助全面理解由权利要求及其等同物限定的本发明的示例性实施例。下述描述包括多种具体细节来帮助理解，但这些细节应认为仅仅是示例性的。因此，本领域普通技术人员应认识到，在不背离本发明的范围和精神的情况下，可以对本文中描述的实施例进行多种改变和修改。此外，为了清楚和简洁起见，省略了公知功能和结构的描述。此外，贯穿附图，相同参考数字用于相似功能和操作。

根据本发明实施例的支持离散数据表示的多层人工神经网络的正向运算，包括两层或者两层以上的多个神经元。对于每一层来说，输入神经元向量首先和权值向量进行点积运算，结果经过激活函数得到输出神经元。其中激活函数可以是sigmoid函数，tanh、relu、softmax函数等非线性函数，支持将激活后的输出神经元离散化表示或连续化表示。在另一种实施例中，激活函数也可以是线性函数。

对于离散数据表示的输入神经元向量或离散数据表示的权值向量的点积运

算，本装置支持将点积运算转换为数据的移位、取非、异或等位运算。对于数据的表示方式，本装置支持数据离散表示或非离散表示，用户可以自定义哪一个层的哪些数据采用离散表示形式或非离散表示，并且可以根据具体需要自定义离散数据的位数，从而代替表示的真实数据的个数，例如设定为 1 比特、2 比特、3 比特等位数的离散数据，分别可以表示 2 个、4 个、8 个真实数据。

图 1 示出了根据本发明实施例的用于执行支持离散数据表示的人工神经网络正向运算的装置的整体结构的示例框图。如图 1 所示，该装置包括指令缓存单元 1、控制器单元 2、数据访问单元 3、互联模块 4、主运算模块 5 和多个从运算模块 6，可选地还包括连续离散转换模块 7。指令缓存单元 1、控制器单元 2、数据访问单元 3、互联模块 4、主运算模块 5 和从运算模块 6、连续离散转换模块 7 均可以通过硬件电路（例如包括但不限于 FPGA、CGRA、专用集成电路 ASIC、模拟电路和忆阻器等）实现。特别的，本装置可以对离散数据提供存储和运算支持。

指令缓存单元 1 通过数据访问单元 3 读入指令并缓存读入的指令。

控制器单元 2 从指令缓存单元 1 中读取指令，将指令译成控制其他模块行为的微指令，所述其他模块例如数据访问单元 3、主运算模块 5 和从运算模块 6 等。

数据访问单元 3 能够访存外部地址空间，直接向装置内部的各个缓存单元读写数据，完成数据的加载和存储。该数据是离散表示的或非离散表示的。该单元用来设计可以读取离散表示的数据。

互联模块 4 用于连接主运算模块和从运算模块，可以实现成不同的互连拓扑（如树状结构、环状结构、网格状结构、分级互连、总线结构等）

图 2 示意性示出了互联模块 4 的一种实施方式：H 树模块。H 树模块 4 构成主运算模块 5 和多个从运算模块 6 之间的数据通路，并具有 H 树的结构。H 树是由多个节点构成的二叉树通路，每个节点将上游的数据同样地发给下游的两个节点，将下游的两个节点返回的数据进行合并，并返回给上游的节点。例如，在每层人工神经网络开始计算阶段，主运算模块 5 内的神经元数据该数据可以是离散表示或非离散表示的通过 H 树模块 4 发送给各个从运算模块 6；当从运算模块 6 的计算过程完成后，每个从运算模块输出的神经元的值会在 H 树中逐级拼成一个完整的由神经元组成的向量，作为中间结果向量。针对于离散数据表示的运算，

我们特别提到了在主从运算模块内部的专用于离散数据运算的运算模块见图 7。以神经网络全连接层进行说明，假设装置中共有 N 个从运算模块，则中间结果向量按 N 分段，每段有 N 个元素，第 i 个从运算模块计算每段中的第 i 个元素。N 个元素经过 H 树模块拼成长度为 N 的向量并返回给主运算模块。所以如果网络只有 N 个输出神经元，则每个从运算单元只需输出单个神经元的值，若网络有 m*N 个输出神经元，则每个从运算单元需输出 m 个神经元值。H 树模块在存储和传输数据的过程中均支持离散数据表示。

图 3 示出了根据本发明实施例的用于执行人工神经网络正向运算的装置中主运算模块 5 的结构的示例框图。如图 3 所示，主运算模块 5 包括运算单元 51、数据依赖关系判断单元 52 和支持离散数据表示的神经元缓存单元 53。

支持离散数据表示的神经元缓存单元 53 用于缓存主运算模块 5 在计算过程中用到的输入数据和输出数据。

运算单元 51 完成主运算模块 5 的各种运算功能。对于运算因子全是离散数据的情况，可以通过查表实现离散数据与离散数据的加减乘除运算。例如 2 位的离散数据，可以表示 4 个连续数据值。对于 4 个连续数据共有 4*4=16 种组合。对于每种加减乘除运算的操作，可以制作并维护该 4*4 的索引表，通过索引表找到对应的计算值。4 种运算共需要 4 张 4*4 的索引表。

对于运算因子包含离散数据和连续数据的情况，可以针对不同离散数据，为加、减、乘、除运算预先设定相应的位操作。例如，可以采取按位异或后乘 2 的相应位次幂之后累加求和的方式代替离散数据与连续数据的点积运算。例如，对于乘法操作，乘法因子数据如果存在离散表示的，可以通过离散数据索引相应的操作（例如，对相应数据的按位异或、取非、移位等操作）代替和该离散数据表示的连续数据的乘法操作，从而减少了乘法器部件数量。例如对于连续数据与离散数据的乘法操作，-1/2 乘以 16。传统的乘法器部件会将-1/2 与 16 直接做乘法。在运算单元 51 中，由于离散数据的可能性较少，可以通过查找索引这样一种开关判断的方法代替了运算单元的功能。例如，可以规定-1/2 的离散数据表示方法为 01。如果一个运算因子是-1/2，则运算单元 51 接收到的离散数据为 01。运算单元 51 便采用离散数据 01 对应的操作。通过对于 16 的 8 位定点数表示 00010000 符号位取反，向右移 1 位得到 10001000，十进制表示为-8。对于除法操作，16

除以-2。其中 16 是连续数据，-2 是离散数据。如果规定离散数据-2 二进制表示为 10。运算单元便采用离散数据 10 对应的除法操作。通过对 16 的 8 位定点数表示 0001000 右移 1 位之后符号位取反得到 10001000，十进制表示为-8 得到结果。加法和减法操作与上述过程类似。根据离散数据的二进制作为一个索引，索引到按位左移、右移、异或等操作。经过该操作后实现了与离散数据表示的真实数据的相加或者相减操作。

依赖关系判断单元 52 是运算单元 51 读写神经元缓存单元 53 的端口，同时能够保证神经元缓存单元中数据的读写一致性。同时，数据依赖关系判断单元 52 也负责将读取数据通过互联模块 4 发送给从运算模块，而从运算模块 6 的输出数据通过互联模块 4 直接发送给运算单元 51。控制器单元 2 输出的指令发送给计算单元 51 和数据依赖关系判断单元 52，来控制其行为。

图 4 示出了根据本发明实施例的用于执行支持离散数据表示的人工神经网络正向运算的装置中从运算模块 6 的结构的示例框图。如图 4 所示，每个从运算模块 6 包括运算单元 61、数据依赖关系判定单元 62、支持离散数据表示的神经元缓存单元 63 和支持离散数据表示的权值缓存单元 64。

运算单元 61 接收控制器单元 2 发出的微指令并进行算数逻辑运算。对于运算因子全是离散数据的情况，可以通过查表实现离散数据与离散数据的加减乘除运算。例如 2 位的离散数据，可以表示 4 个连续数据值。对于 4 个连续数据共有 4*4=16 种组合。对于每种加减乘除运算的操作，可以制作并维护该 4*4 的索引表，通过索引表找到对应的计算值。4 种运算共需要 4 张 4*4 的索引表。

对于运算因子包含离散数据和连续数据的情况，可以针对不同离散数据，为加、减、乘、除运算预先设定相应的位操作。例如，可以采取按位异或后乘 2 的相应位次幂之后累加求和的方式代替离散数据与连续数据的点积运算。例如，对于乘法操作，乘法因子数据如果存在离散表示的，可以通过离散数据索引相应的操作（例如，对相应数据的按位异或、取非、移位等操作）代替和该离散数据表示的连续数据的乘法操作，从而减少了乘法器部件数量。例如对于连续数据与离散数据的乘法操作，-1/2 乘以 16。传统的乘法器部件会将-1/2 与 16 直接做乘法。在运算单元 51 中，由于离散数据的可能性较少，可以通过查找索引这样一种开关判断的方法代替了运算单元的功能。例如，可以规定-1/2 的离散数据表示方法

为 01。如果一个运算因子是-1/2，则运算单元 51 接收到的离散数据为 01。运算单元 51 便采用离散数据 01 对应的操作。通过对于 16 的 8 位定点数表示 00010000 符号位取反，向右移 1 位得到 10001000，十进制表示为-8。对于除法操作，16 除以-2。其中 16 是连续数据，-2 是离散数据。如果规定离散数据-2 二进制表示为 10。运算单元便采用离散数据 10 对应的除法操作。通过对 16 的 8 位定点数表示 0001000 右移 1 位之后符号位取反得到 10001000，十进制表示为-8 得到结果。加法和减法操作与上述过程类似。根据离散数据的二进制作为一个索引，索引到按位左移、右移、异或等操作。经过该操作后实现了与离散数据表示的真实数据的相加或者相减操作。

数据依赖关系判断单元 62 负责计算过程中对神经元缓存单元的读写操作。数据依赖关系判断单元 62 执行读写操作之前会首先保证指令之间所用的数据不存在读写一致性冲突。例如，所有发往数据依赖关系单元 62 的微指令都会被存入数据依赖关系单元 62 内部的指令队列里，在该队列中，读指令的读取数据的范围如果与队列位置靠前的写指令写数据的范围发生冲突，则该指令必须等到所依赖的写指令被执行后才能够执行。

支持离散数据表示的神经元缓存单元 63 缓存该从运算模块 6 的输入神经元向量数据和输出神经元值数据。该数据可以以离散数据的形式存储和传输。

支持离散数据表示的权值缓存单元 64 缓存该从运算模块 6 在计算过程中需要的权值数据。该数据根据用户定义可以是离散表示的或不是。对于每一个从运算模块 6，都只会存储全部输入神经元与部分输出神经元之间的权值。以全连接层为例，输出神经元按照从运算单元的个数 N 进行分段，每段的第 n 个输出神经元对应的权值存放在第 n 个从运算单元中。

从运算模块 6 实现每层人工神经网络正向运算过程中可以并行的前半部分。该模块中的数据存储以及运算都支持离散数据表示。以人工神经网络全连接层（MLP）为例，过程为 y=f（wx+b），其中权值矩阵 w 和输入神经元向量 x 的乘法可以划分为不相关的并行计算子任务，out 与 in 是列向量，每个从运算模块 6 只计算 in 中相应的部分标量元素与权值矩阵 w 对应的列的乘积，得到的每个输出向量都是最终结果的一个待累加的部分和，这些部分和在互联模块 4 中逐级两两相加得到最后的结果。这个结果可以是离散数据表示的。所以计算过程变成了

并行的计算部分和的过程和后面的累加的过程。每个从运算模块 6 计算出输出神经元值，所有的输出神经元值在互联模块 4 中拼成得到中间结果向量。每个从运算模块 6 只需要计算出中间结果向量 y 中与本模块对应的输出神经元值即可。互联模块 4 对所有从运算模块 6 输出的神经元值求和，得到最终的中间结果向量 y。主运算模块 5 基于中间结果向量 y 进行后续计算，比如加偏置、池化（例如最大值池化（MAXPOOLING）或平均值池化（AVGPOOLING）等）、做激活和做采样等。

图 8 示出了运算单元的结构框图，其可用于主运算模块中的运算单元 51 或从运算模块中的运算单元 61。运算过程中输入数据可以是离散数据或连续数据。数据类型判断单元 71 判断输入数据全是连续数据、全是离散数据或是既包含连续数据又包含离散数据的混合数据。当输入数据全是连续数据时，连续数据运算单元 72 执行相应运算。

当输入数据全是离散数据时，离散数据运算单元 73 执行相应运算。对于运算因子全是离散数据的情况，可以通过查表实现离散数据与离散数据的加减乘除运算。例如 2 位的离散数据，可以表示 4 个连续数据值。对于 4 个连续数据共有 4*4=16 种组合。对于每种加减乘除运算的操作，我们制作并维护该 4*4 的索引表，通过索引表找到对应的计算值。4 种运算共需要 4 张 4*4 的索引表。

当输入数据是混合数据时，运算决定单元 74 根据其中的离散数据决定应对其执行何种操作。可以针对不同的离散数据分别预先设置相应操作。然后，混合数据运算单元 75 根据运算决定单元 74 的决定结果，执行相应操作。对于运算因子包含离散数据和连续数据的情况，可以针对不同离散数据，为加、减、乘、除运算预先设定相应的位操作。例如，可以采取按位异或后乘 2 的相应位次幂之后累加求和的方式代替离散数据与连续数据的点积运算。例如，对于乘法操作，乘法因子数据如果存在离散表示的，可以通过离散数据索引相应的操作（例如，对相应数据的按位异或、取非、移位等操作）代替和该离散数据表示的连续数据的乘法操作，从而减少了乘法器部件数量。例如对于连续数据与离散数据的乘法操作，-1/2 乘以 16。传统的乘法器部件会将-1/2 与 16 直接做乘法。在运算单元 51 中，由于离散数据的可能性较少，可以通过查找索引这样一种开关判断的方法代替了运算单元的功能。例如，可以规定-1/2 的离散数据表示方法为 01。如果一个

运算因子是-1/2，则运算单元 51 接收到的离散数据为 01。运算单元 51 便采用离散数据 01 对应的操作。通过对于 16 的 8 位定点数表示 00010000 符号位取反，向右移 1 位得到 10001000，十进制表示为-8。对于除法操作，16 除以-2。其中 16 是连续数据，-2 是离散数据。如果规定离散数据-2 二进制表示为 10。运算单元便采用离散数据 10 对应的除法操作。通过对 16 的 8 位定点数表示 0001000 右移 1 位之后符号位取反得到 10001000，十进制表示为-8 得到结果。加法和减法操作与上述过程类似。根据离散数据的二进制作为一个索引，索引到按位左移、右移、异或等操作。经过该操作后实现了与离散数据表示的真实数据的相加或者相减操作。

图 9 示出了连续离散转换单元。用户可以定义采用该模块将连续数据转换为离散数据或不采用。输入连续数据，输出离散数据。该单元包括随机数产生模块、判断模块、运算模块。对于输入的连续数据通过运算模块得到运算后的结果，经由判断模块用随机数与运算后的结果比较，判断随机数落在哪一个区间，从而决定出输出的离散数据的具体值。例如用户定义产生二元离散数据。对于输入的任意连续数据 x。经由运算模块计算出结果 y,y=abs(clip(-1,1))。之后通过判断模块，如果随机数大于 y，则输出的离散数据是 1，反之输出的离散数据是 0。离散数据 1 和 0 分别代表了连续数据的-1 和+1。将得到的离散数据存储回内存中。等待主从运算模块中的运算单元使用，产生相应的操作。

正向过程中的权值数据、输出输入数据可以采用离散数据表示或不采用。对于连续数据的乘法操作，可以通过基于离散数据的异或、取非、位移等方式代替连续数据的乘法操作。例如权值用 1 比特离散数据表示，0 代表+1,1 代表-1，通过对与权值相乘数据的符号位异或，实现了对权值的乘法运算。

根据本发明实施例，还提供了在前述装置上执行人工神经网络正向运算的指令集。指令集中包括 CONFIG 指令、COMPUTE 指令、IO 指令、NOP 指令、JUMP 指令和 MOVE 指令等，其中：

CONFIG 指令在每层人工神经网络计算开始前配置当前层计算需要的各种常数；

COMPUTE 指令完成每层人工神经网络的算术逻辑计算；

IO 指令实现从外部地址空间读入计算需要的输入数据以及在计算完成后将

数据存回至外部空间，该数据支持离散化表示；

NOP 指令负责清空当前装置内部所有微指令缓存队列中的微指令，保证 NOP 指令之前的所有指令全部指令完毕。NOP 指令本身不包含任何操作；

JUMP 指令负责控制器将要从指令缓存单元读取的下一条指令地址的跳转，用来实现控制流的跳转；

MOVE 指令负责将装置内部地址空间某一地址的数据搬运至装置内部地址空间的另一地址，该过程独立于运算单元，在执行过程中不占用运算单元的资源。

图 5 示出了根据本发明实施例的神经网络正向运算过程的示例框图。在不同从运算模块 6 中，输入神经元向量分别与该从运算模块 6 的权值向量进行点积运算，得到对应的输出神经元值，所有这些输出神经元值组成中间结果向量，该中间结果向量经过加偏置向量以及激活运算得到该层神经网络的最终输出神经元向量，公式描述为 out=f（w*in+b），其中 out 输出神经元向量、in 是输入神经元向量、b 是偏置向量，w 是权值矩阵，f 是激活函数。每个从运算模块 6 的权值向量是权值矩阵中与该从运算模块 6 相对应的列向量。互联模块将输入神经元向量[in0,…,inN]发送给所有的从运算单元，暂存在神经元缓存单元中。对于第 i 个从运算单元，计算其相应的权值向量[w_i0,…,w_iN]与输入神经元向量的点积。从运算单元输出的结果经过互联模块拼成完整的输出向量并返回给主运算单元，在主运算单元中进行激活运算，得到最后的输出神经元向量[out0,out1,out2,…,outN]。

图6是示出根据一个实施例的单层支持离散数据表示的人工神经网络正向计算的一种实施方法。该流程图描述利用本发明的装置和指令集实现图5所示的一种单层离散数据表示的人工神经网络正向运算过程。

步骤S1.1，将初始指令存放到指令存储单元1中；

步骤S1.2，从指令存储单元1中读取一条指令；

步骤S1.3，对上述指令进行译码；

步骤S1.4，根据译码得到的控制信号，进行相应操作；

步骤S1.5，将操作结果写回到相应存储中。

在步骤S1.1中，可以存入初始化IO指令，用于搬运后续指令。

在步骤S1.2中，可读取的指令包括但不限于CONFIG指令、COMPUTE指令、IO指令、NOP指令、JUMP指令和MOVE指令等。

在步骤S1.3中，根据指令的操作类型（CONFIG，COMPUTE，IO，NOP，JUMP，MOVE等）译码得到相应模块的控制信号。对于CONFIG指令，译码得到配置其余模块的配置信息。对于COMPUTE指令，译码得到主从运算模块的控制信号，控制不同离散数据采取的对应操作。对于IO指令，译码得到数据访问模块的控制信号。对于NOP指令，不产生实际控制信号，只用于清空当前装置内部所有控制信号缓存队列中的控制信号，保证NOP指令之前的所有指令全部执行完毕。对于JUMP指令，得到跳转指令流的控制信号。对于MOVE指令，得到在装置内部搬运数据的控制信号。

在步骤S1.4中，上述模块2-6根据控制信号执行相应操作。以执行支持离散数据表示的神经网络正向的COMPUTE指令为例，互连模块将输入神经元向量[in0,…,inN]发送给所有的从运算模块，暂存在神经元缓存单元中。对于第i个从运算模块，计算其相应的权值向量[w_i0,…,w_iN]与输入神经元向量的点积。从运算模块输出的结果经过互连模块拼成完整的输出向量并返回给主运算模块，在主运算模块中进行激活运算，得到最后的输出神经元向量[out0,out1,out2,…,outN]。

在步骤S1.5中，各个模块将操作结果写回到相应缓存中。以执行离散数据表示的神经网络正向的运算为例，主运算模块得到的输出神经元向量被写回到存储单元。

图7是示出根据一个实施例的单层人工神经网络正向运算的另一种更详细的实施方法。该流程图描述利用本发明的装置和指令集实现图4所示的一种单层神经网络正向运算的过程。

在步骤S1，在指令缓存单元1的首地址处预先存入一条IO指令。

在步骤S2，运算开始，控制器单元2从指令缓存单元1的首地址读取该条IO指令，根据译出的微指令，数据访问单元3从外部地址空间读取相应的所有人工神经网络运算指令，并将其缓存在指令缓存单元1中。

在步骤S3，控制器单元2接着从指令缓存单元读入下一条IO指令，根据译出的微指令，数据访问单元3从外部地址空间读取主运算模块5需要的所有数据（例如，包括输入神经元向量、插值表、常数表和偏置等）至主运算模块5的神经元缓存单元53，该数据支持离散表示，可以是全部离散或部分离散。

在步骤S4，控制器单元2接着从指令缓存单元读入下一条IO指令，根据译出

的微指令，数据访问单元3从外部地址空间读取从运算模块6需要的权值矩阵数据，该数据支持离散表示，可以是全部离散或部分离散。

在步骤S5，控制器单元2接着从指令缓存单元读入下一条CONFIG指令，根据译出的微指令，装置配置该层神经网络计算需要的各种常数。例如，运算单元51、61根据微指令里的参数配置单元内部寄存器的值，所述参数例如包括本层计算的精度设置、激活函数的数据（例如本层计算的精度位，Lrn层算法的rang参数，AveragePooling层算法窗口大小的倒数等）。

在步骤S6，控制器单元2接着从指令缓存单元读入下一条COMPUTE指令，根据译出的微指令，主运算模块5首先通过互联模块4将输入神经元向量发给各从运算模块6，保存至从运算模块6的神经元缓存单元63。

在步骤S7，根据COMPUTE指令译出的微指令，从运算模块6的运算单元61从权值缓存单元64读取权值向量（权值矩阵中对应于该从运算模块6的列向量）,从神经元缓存单元读取输入神经元向量，完成权值向量和输入神经元向量的点积运算，将中间结果通过互联返回，对于离散数据，自定义采用异或等位运算代替点积运算或不采用。例如对于1比特的离散数据表示，0代表+1,1代表-1，通过对与权值相乘数据的符号位异或，实现了对权值的乘法运算。。

在步骤S8，在互联模块4中，各从运算模块6返回的中间结果被逐级拼成完整的中间结果向量。

在步骤S9，主运算模块5得到互联模块4的返回值，根据COMPUTE指令译出的微指令，从神经元缓存单元53读取偏置向量，与互联模块4返回的向量相加，然后再对相加结果做激活，该装置支持用户自定义是否将激活后的结果离散化表示。并将最后的输出神经元向量写回至神经元缓存单元53。

在步骤S10，控制器单元接着从指令缓存单元读入下一条IO指令，根据译出的微指令，数据访问单元3将神经元缓存单元53中的输出神经元向量存至外部地址空间指定地址，运算结束。

对于人工神经网络批归一化运算(Batch Normalization)运算步骤与上述过程相仿。通过提供的指令集，控制器完成以下过程。控制器控制数据访问单元读入输入的数据，之后控制主从运算模块根据 batch 大小求出各自位置的均值以及方差或使用设定好的均值方差。之后控制器控制对应位置的输入数据减去均值除以

方差。最后控制器控制用处理后的数据与学习参数相乘后加上另一个学习参数。

对于多层人工神经网络，其实现过程与单层神经网络类似，当上一层人工神经网络执行完毕后，下一层的运算指令会将主运算单元中存储的上一层的输出神经元地址作为本层的输入神经元地址。同样地，指令中的权值地址和偏置地址也会变更至本层对应的地址。

通过采用用于执行人工神经网络正向运算的装置和指令集，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正向运算的支持。

通过采用针对多层人工神经网络正向运算的专用片上缓存，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正向运算性能瓶颈的问题。

通过采用离散数据表示的方法，相较于浮点数、定点数等表示方法，大大较少了装置的存储能耗等开销。可以再有限的面积上优化结构布局，提高运算速度或性能能耗比等指标。

前面的附图中所描绘的进程或方法可通过包括硬件（例如，电路、专用逻辑等）、固件、软件（例如，被具体化在非瞬态计算机可读介质上的软件），或两者的组合的处理逻辑来执行。虽然上文按照某些顺序操作描述了进程或方法，但是，应该理解，所描述的某些操作能以不同顺序来执行。此外，可并行地而非顺序地执行一些操作。并且对于离散数据的表示问题，应该理解可以选择哪些数据离散化表示、哪些连续表示。数据是否离散表示的精神贯穿于整个运算过程中。

在前述的说明书中，参考其特定示例性实施例描述了本发明的各实施例。显然，可对各实施例做出各种修改，而不背离所附权利要求所述的本发明的更广泛的精神和范围。相应地，说明书和附图应当被认为是说明性的，而不是限制性的。

权 利 要 求

1、一种支持离散数据表示的用于执行人工神经网络正向运算的装置，包括指令缓存单元、控制器单元、数据访问单元、互联模块、主运算模块、多个从运算模块，其中：

指令缓存单元用于通过数据访问单元读入指令并缓存读入的指令；

控制器单元用于从指令缓存单元读取指令，并将该指令译码成控制互联模块、主运算模块、以及从运算模块行为的微指令；

数据访问单元用于从外部地址空间向主运算模块和各从运算模块的相应数据缓存单元中写入离散数据或连续数据或从所述数据缓存单元向外部地址空间读离散数据或连续数据；

每层神经网络开始正向计算的阶段，主运算模块通过互联模块向所有的从运算模块传输本层的离散或连续的输入神经元向量，在从运算模块的计算过程完成后，互联模块逐级将各从运算模块的离散或连续化的输出神经元值拼成中间结果向量，其中，当输入数据是离散数据与连续数据的混合数据时，从运算模块针对不同离散数据采取预先设置的相应计算方式；

主运算模块用于利用中间结果向量完成后续计算，当输入数据是离散数据与连续数据的混合数据时，主运算模块针对不同离散数据采取预先设置的相应计算方式。

2、根据权利要求 1 所述的装置，其中，离散数据表示指用特定的离散数字代替真实的连续数据的表示方式。

3、根据权利要求 1 所述的装置，其中，多个从运算模块利用相同的离散或连续的输入神经元向量和各自不同的离散或连续的权值向量，并行地计算出各自的离散或连续输出神经元值。

4、根据权利要求 1 所述的装置，其中，主运算模块对中间结果向量执行以下任一项操作：

加偏置操作，在中间结果向量上加上偏置；

对中间结果向量进行激活，激活函数 active 是非线性函数 sigmoid，tanh，relu，softmax 中的任一个或线性函数；

采样操作，将中间结果向量与随机数比较，大于随机数则输出 1，小于随机数则输出 0；或者

池化操作，包括最大值池化或平均值池化。

5、根据权利要求 1 所述的装置，其中，从运算模块包括输入神经元缓存单元，用于缓存离散或者连续的输入神经元向量。

6、根据权利要求 1 所述的装置，其中，互联模块构成主运算模块和所述多个从运算模块之间的连续或离散化数据的数据通路，互

7、根据权利要求 1 所述的装置，其中，主运算模块包括运算单元、数据依赖关系判断单元和神经元缓存单元，其中：

神经元缓存单元用于缓存主运算模块在计算过程中用到的离散或连续表示的输入数据和输出数据；

运算单元完成主运算模块的各种运算功能，当输入数据是离散数据与连续数据的混合数据时，针对不同离散数据采取预先设置的相应计算方式；

数据依赖关系判断单元是运算单元读写神经元缓存单元的端口，保证对神经元缓存单元中连续数据或离散数据读写不存在一致性冲突，并且负责从神经元缓存单元读取输入离散或连续的神经元向量，并通过互联模块发送给从运算模块；以及

来自互联模块的中间结果向量被发送到运算单元。

8、根据权利要求 1 所述的装置，其中，每个从运算模块包括运算单元、数据依赖关系判定单元、神经元缓存单元和权值缓存单元，其中：

运算单元接收控制器单元发出的微指令并进行算数逻辑运算，当输入数据是离散数据与连续数据的混合数据时，针对不同离散数据采取预先设置的相应计算方式；

数据依赖关系判断单元负责计算过程中对支持离散数据表示的神经元缓存单元和支持离散数据表示的权值缓存单元的读写操作，保证对支持离散数据表示的神经元缓存单元和支持离散数据表示的权值缓存单元的读写不存在一致性冲突；

神经元缓存单元缓存输入神经元向量的数据以及该从运算模块计算得到的输出神经元值；以及

权值缓存单元缓存该从运算模块在计算过程中需要的离散或连续表示的权值向量。

9、根据权利要求 7 或 8 所述的装置，其中，数据依赖关系判断单元通过以下方式保证读写不存在一致性冲突：判断尚未执行的微指令与正在执行过程中的微指令的数据之间是否存在依赖关系，如果不存在，允许该条微指令立即发射，否则需要等到该条微指令所依赖的所有微指令全部执行完成后该条微指令才允许被发射。

10、根据权利要求 7 或 8 所述的装置，其中主运算模块或从运算模块中的运算单元包括运算决定单元和混合数据运算单元，当输入数据是混合数据时，运算决定单元根据其中的离散数据决定应对该混合数据执行何种操作，然后，混合数据运算单元根据运算决定单元的决定结果，执行相应操作。

11、根据权利要求 9 所述的装置，其中主运算模块或从运算模块中的所述运算单元还包括离散数据运算单元和连续数据运算单元中的至少一个，以及数据类型判断单元，当输入数据全是离散数据时，由离散数据运算单元根据输入的离散数据通过查表执行相应操作，当输入数据全是连续数据时，由连续数据运算单元执行相应操作。

12、根据权利要求 1 所述的装置，还包括连续离散转换单元，连续离散转换单元包括预处理模块、距离计算模块、和判断模块，假设使用 M 个离散数据，$M=2^m$，$m \geq 1$，令这些离散数据分别对应于预定区间[-zone,zone]内的 M 个数值，其中：

预处理模块对于输入的连续数据 x 使用 clip(-zone,zone)运算进行预处理，得到区间[-zone,zone]内的预处理数据 y，其中，如果 $x \leq -zone$ 则 $y = -zone$，如果 $x \geq zone$ 则 $y = zone$，如果 $-zone < x < zone$，则预处理数据 $y = x$；

距离计算模块计算预处理数据 y 与上述各数值之间的距离；以及

判断模块基于该距离计算并输出离散数据。

13、根据权利要求 12 所述的装置，其特征在于以下任意一项或多项：

预定区间[-zone,zone]是[-1,1]或[-2,2]；

M 个数值的绝对值是 2 的幂的倒数；或者

判断模块执行：

输出与该预处理数据 y 距离最近的数值所对应的离散数据，如果有两个数值与该预处理数据距离相等，则输出二者中任一个所对应的离散数据；或者

计算预处理数据 y 分别到距离最近的两个数值中任一个的归一化概率，将这两个数值中任一个所对应的归一化概率与随机数生成模块生成的(0,1)之间的随机数 z 比较，如果该 z 小于该概率则输出该离散数据，否则输出另一离散数据。

14、一种使用根据权利要求 1-13 中的任一项的装置执行单层人工神经网络正向运算的方法，包括：

数据访问单元从外部地址空间读取与该层人工神经网络正向运算有关的所有人工神经网络运算指令，并将其缓存在指令缓存单元中；

连续离散转换模块从外部地址空间读取该层神经网络需要转换的连续数据转换为离散数据后存储回外部地址空间；

数据访问单元从外部地址空间读取主运算模块需要的与该层人工神经网络正向运算有关的所有离散或连续数据至主运算模块的神经元缓存单元；

数据访问单元从外部地址空间读取从运算模块需要的离散表示或连续表示的权值矩阵数据；

配置该层神经网络正向运算需要的各种离散或连续表示的常数；

主运算模块首先通过互联模块将输入神经元向量发给各从运算模块，保存至从运算模块的支持离散数据表示的神经元缓存单元；

从运算模块的运算单元从权值缓存单元读取权值向量,从从运算模块的神经元缓存单元读取输入神经元向量，对于向量中没有离散数据表示的完成权值向量和输入神经元向量的点积运算，对于向量中有离散数据表示的，通过离散数据运算模块，根据离散数据的值判断相应的位操作代替点积运算，将得到的神经元值通过互联模块返回；

在互联模块中，各从运算模块返回的神经元值被逐级拼成完整的中间结果向量；

主运算模块从主运算模块的神经元缓存单元读取离散表示或连续表示的偏置向量，与互联模块返回的中间结果向量相加，然后再对相加结果做激活，得到输

出神经元向量写回至主运算模块的神经元缓存单元；以及

数据访问单元将主运算模块的神经元缓存单元中的输出神经元向量存至外部地址空间指定地址。

15、一种使用根据权利要求1-13中的任一项的装置执行批归一化运算的方法：

数据访问单元从外部地址空间读取与该批归一化正向运算有关的所有人工神经网络运算指令，并将其缓存在指令缓存单元中；

连续离散转换模块从外部地址空间读取该层神经网络需要转换的连续数据转换为离散数据后存储回外部地址空间。

数据访问单元从外部地址空间读取主运算模块需要的与该层批归一化正向运算有关的所有离散或连续数据至主运算模块的神经元缓存单元；

配置该层批归一化正向运算需要的各种离散或连续表示的常数；

主运算模块首先通过互联模块将输入神经元向量发给各从运算模块，保存至从运算模块的支持离散数据表示的神经元缓存单元；

从运算模块的运算单元从权值缓存单元读取权值向量,从从运算模块的神经元缓存单元读取输入神经元向量，对于输入向量计算在每一个批的尺度下的均值和标准差，将得到的神经元值通过互联模块返回；

在互联模块中，各从运算模块返回的神经元值被逐级拼成完整的中间结果向量；

主运算模块从主运算模块的神经元缓存单元读取离散表示或连续表示输入神经元向量，与互联模块返回的均值结果向量相减，然后再对减结果和标准差结果相除，得到输出神经元向量写回至主运算模块的神经元缓存单元；

数据访问单元将主运算模块的神经元缓存单元中的输出神经元向量存至外部地址空间指定地址。

16、一种执行多层人工神经网络正向运算的方法，包括：

针对每一层，执行根据权利要求 14、15 所述的方法，其中：

当针对上一层人工神经网络执行完毕后，将主运算模块中存储的上一层的输出神经元地址作为本层的输入神经元地址，针对所述本层再次执行根据权利要求 14 或 15 所述的方法。

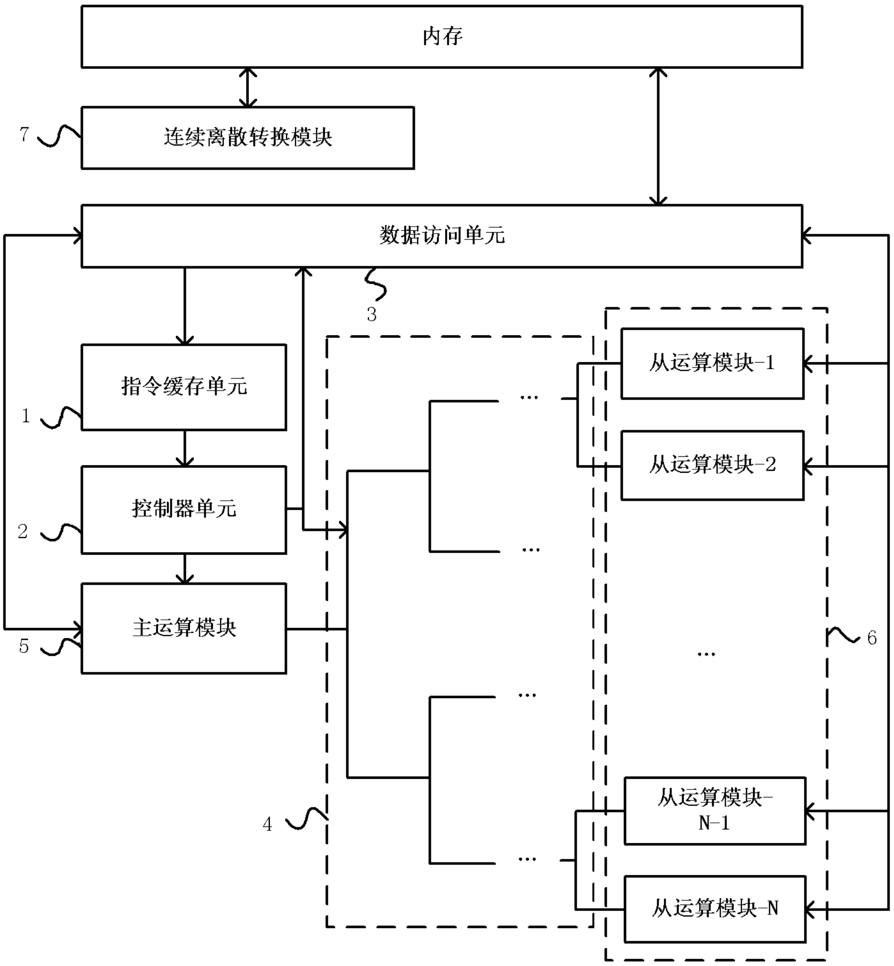

内存
7
连续离散转换模块
数据访问单元
3
指令缓存单元
1
控制器单元
2
主运算模块
5
4
从运算模块-1
从运算模块-2
...
...
...
...
...
从运算模块-N-1
从运算模块-N
6

图 1

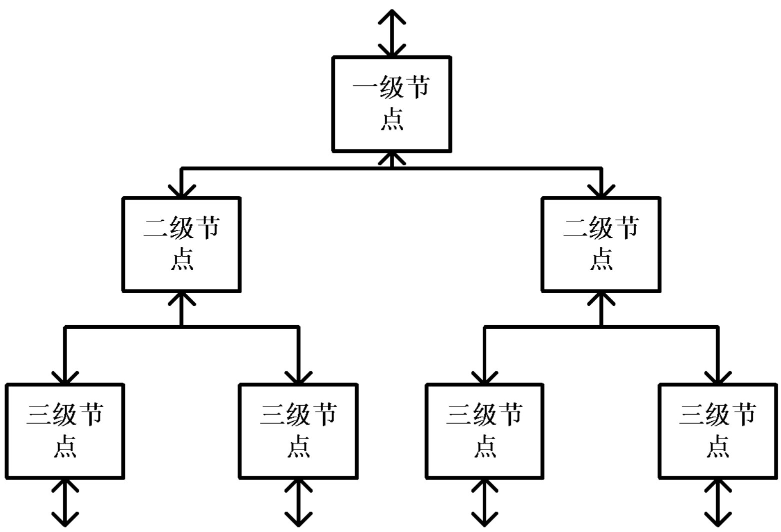

一级节点
二级节点
二级节点
三级节点
三级节点
三级节点
三级节点

图 2

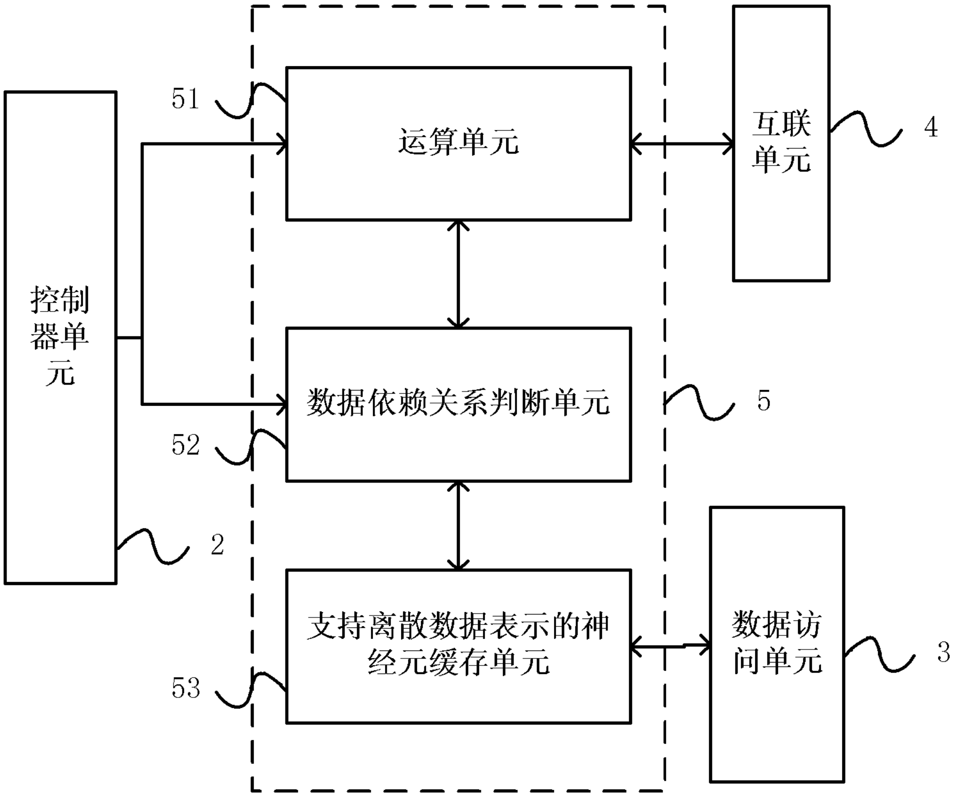

控制器单元
2
51
运算单元
互联单元
4
数据依赖关系判断单元
5
52
支持离散数据表示的神经元缓存单元
53
数据访问单元
3

图 3

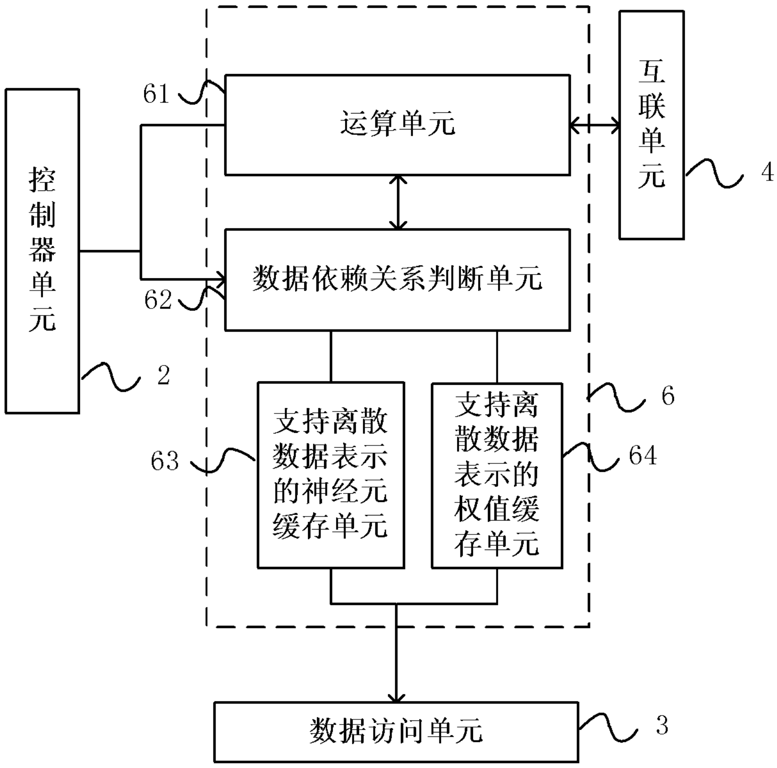

控制器单元
2
61
运算单元
互联单元
4
数据依赖关系判断单元
62
6
支持离散数据表示的神经元缓存单元
63
支持离散数据表示的权值缓存单元
64
数据访问单元
3

图 4

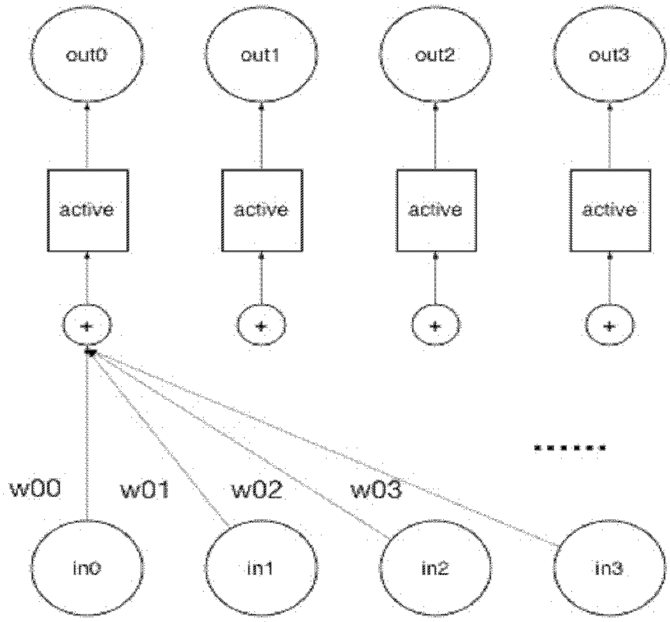

out0
out1
out2
out3
active
active
active
active
+
+
+
+
......
w00
w01
w02
w03
in0
in1
in2
in3

图 5

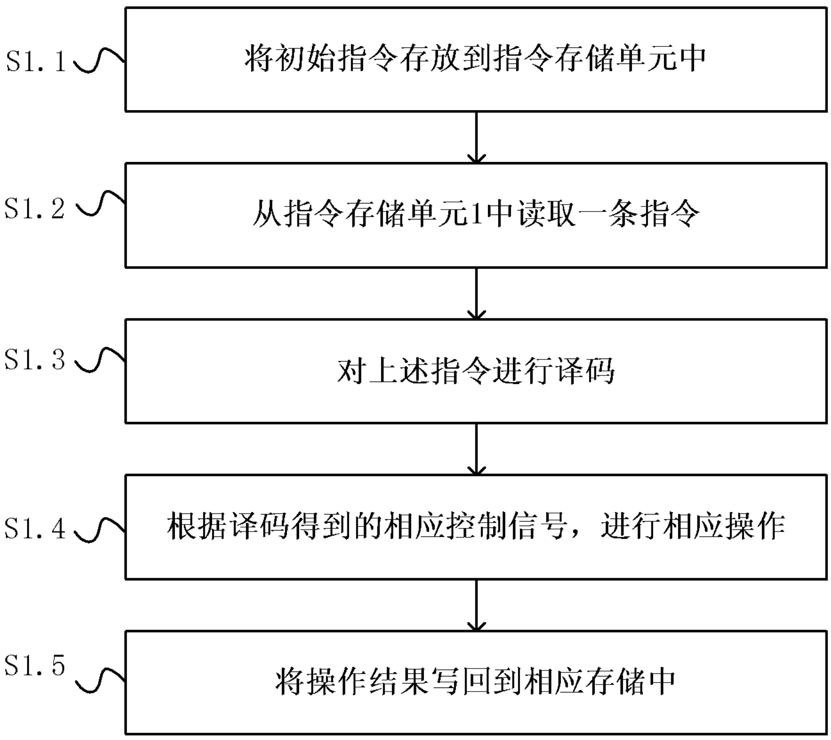

S1.1
将初始指令存放到指令存储单元中
S1.2
从指令存储单元1中读取一条指令
S1.3
对上述指令进行译码
S1.4
根据译码得到的相应控制信号，进行相应操作
S1.5
将操作结果写回到相应存储中

图 6

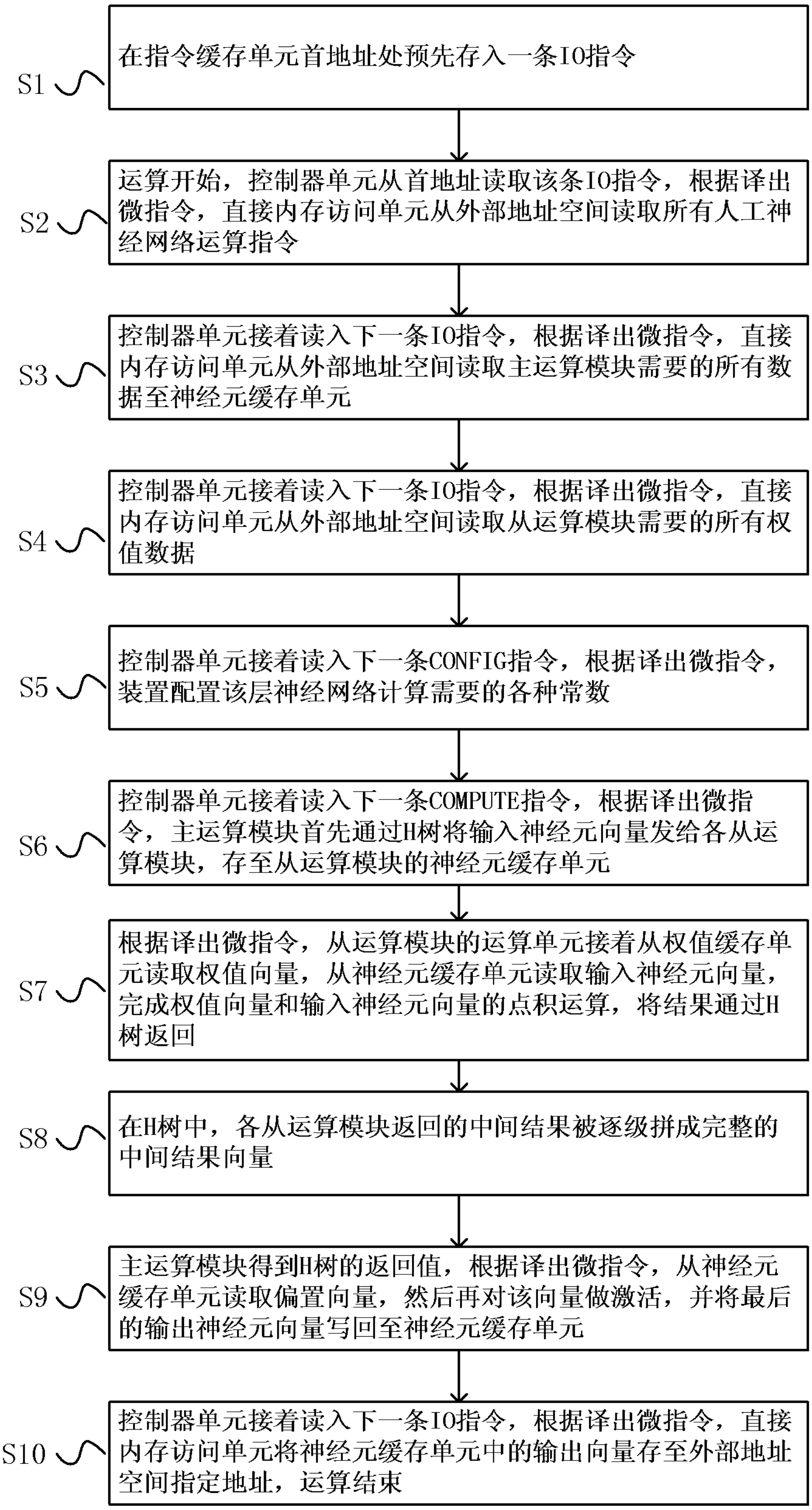
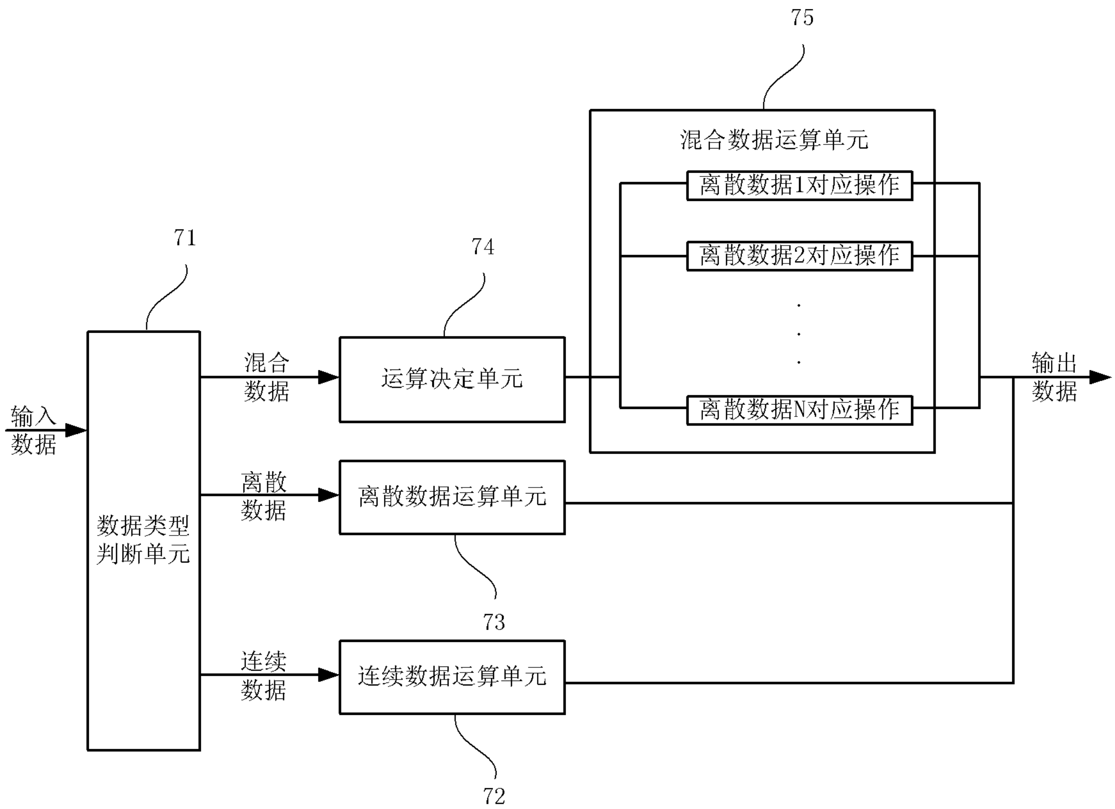
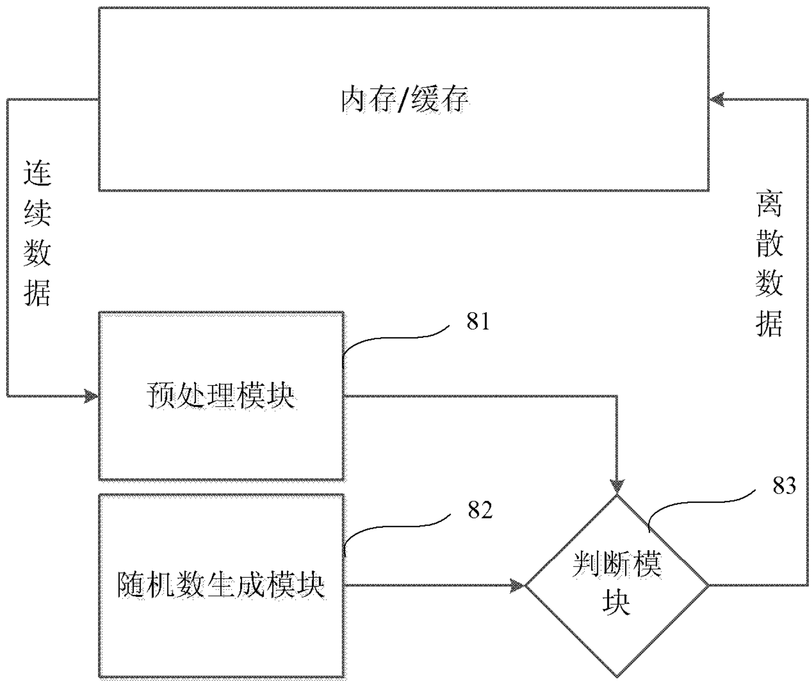

S1
在指令缓存单元首地址处预先存入一条IO指令
S2
运算开始，控制器单元从首地址读取该条IO指令，根据译出微指令，直接内存访问单元从外部地址空间读取所有人工神经网络运算指令
S3
控制器单元接着读入下一条IO指令，根据译出微指令，直接内存访问单元从外部地址空间读取主运算模块需要的所有数据至神经元缓存单元
S4
控制器单元接着读入下一条IO指令，根据译出微指令，直接内存访问单元从外部地址空间读取从运算模块需要的所有权值数据
S5
控制器单元接着读入下一条CONFIG指令，根据译出微指令，装置配置该层神经网络计算需要的各种常数
S6
控制器单元接着读入下一条COMPUTE指令，根据译出微指令，主运算模块首先通过H树将输入神经元向量发给各从运算模块，存至从运算模块的神经元缓存单元
S7
根据译出微指令，从运算模块的运算单元接着从权值缓存单元读取权值向量，从神经元缓存单元读取输入神经元向量，完成权值向量和输入神经元向量的点积运算，将结果通过H树返回
S8
在H树中，各从运算模块返回的中间结果被逐级拼成完整的中间结果向量
S9
主运算模块得到H树的返回值，根据译出微指令，从神经元缓存单元读取偏置向量，然后再对该向量做激活，并将最后的输出神经元向量写回至神经元缓存单元
S10
控制器单元接着读入下一条IO指令，根据译出微指令，直接内存访问单元将神经元缓存单元中的输出向量存至外部地址空间指定地址，运算结束

图 7



75
混合数据运算单元
离散数据1对应操作
离散数据2对应操作
离散数据N对应操作
71
74
混合数据
运算决定单元
输出数据
输入数据
数据类型判断单元
离散数据
离散数据运算单元
73
连续数据
连续数据运算单元
72

图 8



内存/缓存
连续数据
离散数据
81
预处理模块
83
82
判断模块
随机数生成模块

图9

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2016/079431

A. CLASSIFICATION OF SUBJECT MATTER

G06N 3/02 (2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06N; G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CPRSABS; CNTXT; VEN: main, calculation, principal, subordinate, discrete, continuous, data, nerve network, NN, parallel, compute, mixed, controller, buffer, register, access

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 103399486 A (HANGZHOU DIANZI UNIVERSITY) 20 November 2013 (20.11.2013) description, paragraph [0022]	1-10, 12, 13
Y	ZHANG, Xian. "An Algorithm for Training Back-propagation Neural Networks Based on Data Parallelism" China Master's Theses Full-text Database Information and Technology, vol. 2010, no. 05, 10 May 2010 (10.05.2010), ISSN: 1674-0246, chapters 4.1.2, 4.3.2, 4.3.3 and 4.4.1	1-10, 12, 13
A	CN 101963983 A (JIANGSU RIFETUNE SOFTWARE TECHNOLOGY CO., LTD.) 02 February 2011 (02.02.2011) the whole document	1-16
A	CN 103886519 A (BEIJING NORMAL UNIVERSITY) 25 June 2014 (25.06.2014) the whole document	1-16
A	CN 104462459 A (WUHU LERUISI INFORMATION CONSULTING CO., LTD.) 25 March 2015 (25.03.2015) the whole document	1-16

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&"document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
27 October 2016	16 January 2017

Name and mailing address of the ISA
State Intellectual Property Office of the P. R. China
No. 6, Xitucheng Road, Jimenqiao
Haidian District, Beijing 100088, China
Facsimile No. (86-10) 62019451

Authorized officer
CHEN, An'an
Telephone No. (86-10) 62411828

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2016/079431

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 103399486 A	20 November 2013	CN 103399486 B	06 April 2016
CN 101963983 A	02 February 2011	None	
CN 103886519 A	25 June 2014	None	
CN 104462459 A	25 March 2015	None	

国际检索报告

国际申请号

PCT/CN2016/079431

A. 主题的分类

G06N 3/02(2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06N; G06F

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用）)

CPRSABS;CNTXT;VEN:主运算，从运算，离散，连续，数据，神经网络，并行计算，混合，控制器，缓存，访问，main, calculation, principal, subordinate, discrete, continuous, data, nerve network, NN, parallel, compute, mixed, controller, buffer, register, access

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
Y	CN 103399486 A (杭州电子科技大学) 2013年 11月 20日 (2013 - 11 - 20) 说明书第[0022]段	1-10，12-13
Y	张弦. "基于数据并行的BP神经网络训练算法" 《中国优秀硕士学位论文全文数据库信息科技辑》，第2010卷，第05期，2010年 5月 10日 (2010 - 05 - 10), ISSN：1674-0246, 第4.1.2、4.3.2、4.3.3和4.4.1节	1-10，12-13
A	CN 101963983 A (江苏瑞蚨通软件科技有限公司中外合资) 2011年 2月 2日 (2011 - 02 - 02) 全文	1-16
A	CN 103886519 A (北京师范大学) 2014年 6月 25日 (2014 - 06 - 25) 全文	1-16
A	CN 104462459 A (芜湖乐锐思信息咨询有限公司) 2015年 3月 25日 (2015 - 03 - 25) 全文	1-16

☐ 其余文件在C栏的续页中列出。　　☑ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2016年 10月 27日	2017年 1月 16日

ISA/CN的名称和邮寄地址	受权官员
中华人民共和国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088	陈安安
传真号 (86-10)62019451	电话号码 (86-10)62411828

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2016/079431

检索报告引用的专利文件	公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN 103399486 A	2013年 11月 20日	CN 103399486 B	2016年 4月 6日
CN 101963983 A	2011年 2月 2日	无	
CN 103886519 A	2014年 6月 25日	无	
CN 104462459 A	2015年 3月 25日	无	