(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 11 月 2 日 (02.11.2017)

WIPO | PCT

(10) 国际公布号
WO 2017/185347 A1

(51) 国际专利分类号:
G06N 3/063 (2006.01)

(21) 国际申请号: PCT/CN2016/080744

(22) 国际申请日: 2016 年 4 月 29 日 (29.04.2016)

(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: 郭崎 (GUO, Qi); 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)。 陈峋宇 (CHEN, Xunyu); 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, RO, RS, RU,

(54) **Title:** APPARATUS AND METHOD FOR EXECUTING RECURRENT NEURAL NETWORK AND LSTM COMPUTATIONS

(54) **发明名称**: 用于执行循环神经网络和LSTM运算的装置和方法

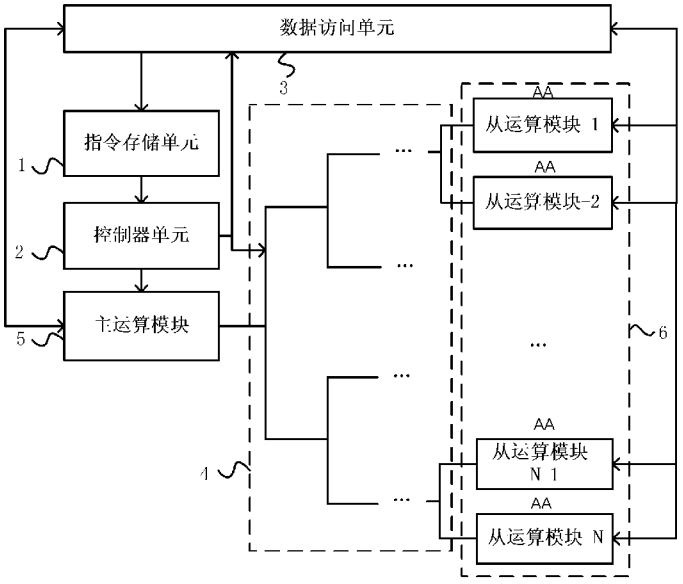

图 1

(57) **Abstract:** An apparatus for executing a recurrent neural network and LSTM, comprising a command storage unit (1), a controller unit (2), a data access unit (3), an interconnection module (4), a main computation module (5), and a plurality of slave computation modules (6). The slave computation modules (6) are used for multiplying input data to obtain a partial sum and saving same until all the data in the neural network is inputted, and returning the results to the main computation module (5); and the main computation module (5) is used for implementing interpolation activation of the sums returned by the slave computation modules (6) in a forward process, and interpolating same in a reverse process to obtain an activation derivative, and multiplying same by a gradient. The present apparatus can solve the problems of insufficient CPU and GPU computational performance and high front end decoding overheads, effectively improving the support for forward computation in multi-layer artificial neural networks.

WO 2017/185347 A1

[见续页]

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))。

(57) 摘要：一种用于执行循环神经网络和LSTM的装置，包括指令存储单元（1）、控制器单元（2）、数据访问单元（3）、互连模块（4）、主运算模块（5）以及多个从运算模块（6）。从运算模块（6）用于将输入数据进行乘加得到部分和并保存，直到神经元数据全都输入将结果返回给主运算模块（5）；主运算模块（5）用于在正向过程时对从运算模块（6）返回的和进行插值激活，在反向过程时插值得到激活导数并且与梯度相乘。该装置能够解决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。

用于执行循环神经网络和 LSTM 运算的装置和方法

技术领域

本发明人工神经网络技术领域，具体涉及循环神经网络（RNN）和 LSTM，特别是一种用于执行循环神经网络和 LSTM 的装置和方法。

背景技术

循环神经网络和 LSTM 被广泛应用于语音识别，语言建模，翻译，图片描述等领域，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。

一种支持循环神经网络和 LSTM 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。该方法的缺点之一是单个通用处理器的运算性能较低，无法满足通常的循环神经网络和 LSTM 运算的性能需求。而多个通用处理器并行执行时，通用处理器之间相互通信又成为了性能瓶颈。另外，通用处理器需要把循环神经网络和LSTM的反向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

另一种支持循环神经网络和 LSTM 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。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，没有对多层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多层人工神经网络运算，带来了大量的额外开销。另外 GPU 只有较小的片上缓存，循环神经网络和 LSTM 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。另外，GPU 只有较小的片上缓存，循环神经网络和 LSTM 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

本发明的一个方面提供了一种用于执行循环神经网络和LSTM的装置，包括指令存储单元、控制器单元、数据访问单元、互连模块、主运算模块、以及多个从运算模块，其中：指令存储单元用于缓存指令；控制器单元用于从指令存储单元读取指令，并将该指令译码成控制互连模块、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内存向主运算模块和各从运算模块的相应数据存储单元中写数据或从所述数据存储单元向内存读数据；互连模块用于，在每层神经网络反向训练开始计算的阶段，主运算模块通过互连模块向所有的从运算模块传输本层的输入梯度向量，在从计算模块的计算过程完成后，互连模块逐级将各从计算模块的输出梯度向量部分和两两相加得到本层的输出梯度向量；从运算模块用于将输入数据进行乘加得到部分和并保存，直到神经元数据全都输入将结果返回给主运算模块；主运算模块用于在正向过程时对从运算模块返回的和进行插值激活，在反向过程时插值得到激活导数并且与梯度相乘。

本发明还提供了一种使用上述装置执行循环神经网络和LSTM运算的方法。

该装置可以应用于以下（包括但不限于）场景中：数据处理、机器人、电脑、打印机、扫描仪、电话、平板电脑、智能终端、手机、行车记录仪、导航仪、传感器、摄像头、云端服务器、相机、摄像机、投影仪、手表、耳机、移动存储、可穿戴设备等各类电子产品；飞机、轮船、车辆等各类交通工具；电视、空调、微波炉、冰箱、电饭煲、加湿器、洗衣机、电灯、燃气灶、油烟机等各类家用电器；以及包括核磁共振仪、B超、心电图仪等各类医疗设备。

附图说明

图1示出了根据本发明实施例的用于执行循环神经网络和LSTM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循环神经网络和LSTM 的装置中互连模块的结构；

图 3 示出了根据本发明实施例的用于执行循环神经网络和 LSTM 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循环神经网络和 LSTM 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循环神经网络和 LSTM 正反向过程的示例框图；

图 6 示出了利用本发明的执行循环神经网络和 LSTM 的装置的运算的过程；

图 7 是循环神经网络的结构；

图 8 是 LSTM 算法的一个 block 的结构；

图 9 示出了本发明的循环神经网络和 LSTM 单层的流程图；

图 10 示出了本发明的循环神经网络和 LSTM 的单层运算的梯度反向传递流程图。

具体实施方式

图 1 示出了本发明实施例的用于执行循环神经网络和 LSTM 运算的装置的整体结构示意图。如图 1 所示，该装置包括指令存储单元 1、控制器单元 2、数据访问单元 3、互连模块 4、主运算模块 5 和多个从运算模块 6。指令存储单元 1、控制器单元 2、数据访问单元 3、互连模块 4、主运算模块 5 和从运算模块 6 均可以通过硬件电路（包括但不限于 FPGA、CGRA、专用集成电路 ASIC、模拟电路和忆阻器）实现。

指令存储单元 1 通过数据访问单元 3 读入指令并缓存读入的指令。指令存储单元 1 可以通过各种不同存储器件（SRAM、DRAM、eDRAM、忆阻器、3D-DRAM 和非易失存储等）实现。

控制器单元 2 从指令存储单元 1 中读取指令，将指令译码成控制其他单元或模块行为的微指令，并发送给所述单元或模块，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 用于访存外部地址空间，直接向装置内部的各个存储单元读写数据，完成数据的加载和存储。

述互连模块用于将所述主运算模块的输入向量分发给所述多个从运算模块，以及将各从运算模块的计算结果合并后返回给主运算模块。图 2 示意性示出了互连模块的结构的一种实施方式的结构示意图。互连模块 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，在该实施例中具有 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。

以典型的循环神经网络和 LSTM 的计算$\text{out} = \sum w \times in_data$为例，主运算模块 5 内的神经元数据通过互连模块 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数据的值会在 H 树中逐级拼成一个完整的由神经元数据组成的向量，作为中间结果向量。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 个元素经过互连模块拼成长度为 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。

在本发明中，主运算模块在正向对从运算模块返回的和进行插值激活，在反向插值得到激活导数并且与梯度相乘。

在本发明中，从运算模块用于将输入数据进行乘加得到部分和并保存，直到神经元全都输入将结果返回给主运算模块。

图 3 示出了根据本发明用于执行循环神经网络和 LSTM 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 和神经元存储单元 53。

神经元存储单元 53 用于缓存主运算模块 5 在计算过程中用到的输入神经元数据和输出神经元数据。运算单元 51 完成主运算模块的各种运算功能。数据依赖关系判断单元 52 是运算单元 51 读写神经元存储单

元 53 的端口，同时能够保证对神经元存储单元 53 中数据的读写不存在一致性冲突。

具体地，数据依赖关系判断单元 52 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。例如，所有发往数据依赖关系单元 52 的微指令都会被存入数据依赖关系单元 52 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。同时，数据依赖关系判断单元 52 也负责从神经元存储单元 53 读取输入梯度向量通过互连模块 4 发送给从运算模块 6，而从运算模块 6 的输出数据通过互连模块 4 直接发送给运算单元 51。控制器单元 2 输出的指令发送给运算单元 51 和依赖关系判断单元 52，来控制其行为。

图 4 示出了根据本发明的用于执行循环神经网络和 LSTM 的装置的从运算模块 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、神经元存储单元 63、权值存储单元 64 和权值梯度存储单元 65。

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。

数据依赖关系判断单元 62 负责计算过程中对存储单元的读写操作。数据依赖关系判断单元 62 保证对存储单元的读写不存在一致性冲突。具体地，数据依赖关系判断单元 62 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。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。

神经元存储单元 63 缓存输入向量数据中与该从运算模块 6 相对应的标量数据以及该从运算模块 6 计算得到的输出向量部分和。

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。对于每一个从运算模块，都只会存储权值矩阵中与该从运算模块 6 所存储的标量数据相对应的列。

权值梯度缓存单元 65 缓存相应从运算模块在更新权值过程中需要的权值梯度数据。每一个从运算模块 6 存储的权值梯度数据与其存储的权值梯度数据相对应。

从运算模块 6 实现循环神经网络和 LSTM 的输出梯度向量的过程中可以并行的前半部分以及权值的更新。

以 $\text{out} = \sum w \times in\_data$ 为例，其中权值矩阵 w 和输入梯度向量 in_data 的乘法可以划分为不相关的并行计算子任务，out 与 in_data 是列向量，每个从运算模块只计算 in_data 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 H 树中逐级两两相加得到最后的结果。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 计算出输出向量的部分和，所有的部分和在互连模块 4 中完成求和运算得到最后的输出向量。每个从运算模块 6 同时将输入向量和正向运算时每层的输出值相乘，计算出权值，以更新本从运算模块 6 存储的权值。正向运算和反向训练是神经网络算法的两个主要过程，神经网络要训练（更新）网络中的权值，首先需要计算输入向量在当前权值构成的网络中的正向输出，这是正向过程，然后根据输出值与输入向量本身的标注值之间的差值，反向逐层训练（更新）每层的权值。在正向计算过程中会保存每一层的输出向量以及激活函数的导数值，这些数据是反向训练过程所需要的，所以在反向训练开始时，这些数据已经保证存在。正向运算中每层的输出值是反向运算开始时已有的数据，可以通过数据访存单元缓存在主运算模块中并通过 H 树发送给从运算模块。主运算模块 5 基于输出梯度向量进行后续计算，例如将输出梯度向量乘以正向运算时的激活函数的导数得到下一层的输入梯度值。正向运算时的激活函

数的导数是在反向运算开始时已有的数据，可以通过数据访存单元缓存在主运算模块中。

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，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；

NOP 指令负责清空当前装至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存储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循环神经网络和 LSTM 正反向过程的示例框图。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 $\text{out} = \sum w \times in\_data$。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。互连模块将输入神经元向量[in0,…,inN]发送给所有的从运算单元，暂存在神经元存储单元中。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连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 显示了利用本发明的装置和指令集实现循环神经网络和 LSTM 运算的过程。

在步骤 S1，在指令存储单元 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 从指令存储单元 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，并将其缓存在指令存储单元 1 中。

在步骤 S3，控制器单元 2 接着从指令存储单元读入下一条 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 的神经元存储单元 53。

在步骤 S4，控制器单元 2 接着从指令存储单元读入下一条 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。

在步骤 S5，控制器单元 2 接着从指令存储单元读入下一条 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）。

在步骤 S6，控制器单元 2 接着从指令存储单元读入下一条 COMPUTE 指令，根据译出的微指令，主运算模块 5 首先通过互连模块 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 的神经元存储单元 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 从权值存储单元 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）,从神经元存储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，将中间结果通过互连模块返回。

在步骤 S8，在互连模块 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 得到互连模块 4 的返回值，根据 COMPUTE 指令译出的微指令，从神经元存储单元 53 读取偏置向量，与互连模块 4 返回的向量相加，然后再对相加结果做激活，并将最后的输出神经元向量写回至神经元存储单元 53。

在步骤 S10，控制器单元接着从指令存储单元读入下一条 IO 指令，根据译出的微指令，数据访问单元 3 将神经元存储单元 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。

图7是循环神经网络的结构。为了解决传统神经网络在时间上对于以前的输入的依赖，正向运算的时候，循环神经网络的输入来自当前时刻的输入以及上一时刻的隐层输出。公式中I为输入数量，H为隐层数量，K为输出数量。其中α_h^t是t时刻第h个输出的中间值，b_h^t是激活后t时刻的第h个输出，δ_h^t表示残差对α_h^t的偏导数，θ 表示激活函数。

正向传播的公式表达为：

$$\alpha_h^t = \sum_{i=1}^{I} w_{ih} x_i^t + \sum_{h'=1}^{H} w_{ih} b_{h'}^{t-1}$$

$$b_h^t = \theta(\alpha_h^t)$$

反向传播的公式表达：

$$\delta_h^t = \theta'(a_h^t)\left(\sum_{k=1}^{K} \delta_k^t w_{hk} + \sum_{h'=1}^{H} \delta_{h'}^{t+1} w_{hh'}\right)$$

其中。通过对于上一时刻的隐层和本层输出的连接来达到综合时间序列的作用。但是这样的循环神经网络存在时间衰减的问题。

图 8 是 LSTM 算法的一个 block 的结构。相对于传统循环神经网络，LSTM 引入了一个 cell 来记录当前时间点的信息。可以看出在 LSTM 算法里，一个 block 由三个门和一个 cell 组成，输入门、输出门、忘记门。

LSTM 算法的主要思想是利用 cell 来记录当前时间的状态，对上一时刻传入 cell 值来达到在不同时间直接传递信息的功能。用输入门和忘记门来控制 cell 的输出里对于当前时间输入和上一时间 cell 的权重。用输出门来控制 cell 的输出。在输入门和忘记门的控制下，合适的信息将会被保存很长时间，一直记录在 cell 里面，这样就解决了循环神经网络随着时间衰减的问题。

图 9 示出了本发明的循环神经网络和 LSTM 单层的流程图。

在步骤 A1，计算出输入门对应的当前时刻输入与权值的乘积和缓存在神经元缓存区，再计算上一时刻单元状态与对应权值的乘积和以及上一时刻隐层与对应权值乘积和都存在缓存区。最后将他们三个相加并且激活得到输入门值。

在步骤 A2，计算出忘记门对应的当前时刻输入与权值的乘积和缓存在神经元缓存区，再计算上一时刻单元状态与对应权值的乘积和以及上一时刻隐层与对应权值乘积和都存在缓存区。最后将他们三个相加并且激活得到忘记门值。

在步骤 A3，计算出输入门对应的当前时刻输入与权值的乘积和缓存在神经元缓存区，再计算上一时刻隐层与对应权值乘积和都存在缓存区。最后将他们两个个相加并且激活得到单元状态中间值缓存到神经元缓存区。然后让中间值和输入门对应相乘，缓存在运算单元（图 7 的 51）缓存区内，再让上一时刻的单元状态和忘记门对应相乘，在运算单元与上一次缓存对应相加，得到单元状态值。

在步骤 A4，计算出输出门对应的当前时刻输入与权值的乘积和缓存在神经元缓存区，再计算当前时刻单元状态与对应权值的乘积和以及上一时刻隐层与对应权值乘积和都存在缓存区。最后将他们三个相加并且激活，得到输出门值。

在步骤 A5，单元状态和输出门对应相乘得到本层输出。

图 10 示出了本发明的循环神经网络和 LSTM 的单层运算的梯度反向传递流程图。

在步骤 B1，计算出下一时刻隐层梯度和对应位置的权值相乘相加加上本层残差和对应权值相乘之和，得到本层输出梯度。

在步骤 B2，让输出梯度和 cell 激活值对应乘加，缓存在神经元缓存区最后乘以激活函数导数得到输出门梯度。

在步骤 B3 将 cell 的状态梯度由当前输出梯度乘上当前输出门值和状态激活的导数存入神经元，再计算上下一时刻 cell 的梯度，输入门和忘记门的梯度以及本时刻输出门梯度乘以对应权值都存入神经元缓存，最后相加得到单元状态梯度。cell 中间值的梯度由当前时刻输入门激活值，cell 激活函数导数以及 cell 状态梯度相乘得到。

在步骤 B4，将当前时刻所有 cell 的状态梯度对应乘以上一时刻 cell 状态输出相加最后与忘记门导数相乘得到忘记门的梯度。

在步骤 B5，将当前时刻所有 cell 的状态梯度对应乘以本时刻 cell 中间值的激活值输出相加最后与输入门导数相乘得到输入门的梯度。

值得一提的是，传统循环神经网络算法应用在此装置上是极大简化了的 LSTM 算法，计算输出的时候只依赖当前时刻输入和上一时刻输出，正反向表达与 LSTM 的运行子过程类似，在此不做赘述。

对于一次完整的循环神经网络和 LSTM 算法，其实现过程与以上典型的计算类似，按照公式取出相应的权值和数据进行加权求和，长时间的时候，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。

通过采用用于执行循环神经网络和 LSTM 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。

通过采用针对循环神经网络和 LSTM 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。

权利要求

1、一种用于执行循环神经网络和 LSTM 运算的装置，包括指令存储单元、控制器单元、互连模块、主运算模块和多个从运算模块，其中，

所述指令存储单元通过数据访问单元读入指令并缓存读入的指令；

所述控制器单元从指令存储单元中读取指令，将指令译码成控制其他单元或模块行为的微指令，然后将各自的微指令分发至各个单元或模块；

所述数据访问单元用于访问外部地址空间，完成数据的加载和存储；

所述互连模块用于将所述主运算模块的输入向量分发给所述多个从运算模块，以及将各从运算模块的计算结果合并后返回给主运算模块；

所述从运算模块用于将输入数据进行乘加得到部分和并保存，直到神经元数据全都输入将结果返回给主运算模块；

所述主运算模块用于在正向过程时对从运算模块返回的和进行插值激活，在反向过程时插值得到激活导数并且与梯度相乘。

2、如权利要求 1 所述的用于执行循环神经网络和 LSTM 运算的装置，所述主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，其中，

所述运算单元用于接收控制器单元发出的微指令并进行算数逻辑运算；

所述数据依赖关系判断单元用于对神经元缓存单元进行读写操作，保证指令之间所用的数据不存在读写一致性冲突；

所述神经元缓存单元用于缓存输入神经元数据和输出神经元数据。

3、如权利要求 1 所述的用于执行循环神经网络和 LSTM 运算的装置，所述从运算模块包括运算单元、数据依赖关系判定单元、神经元存储单元、权值存储单元和权值梯度存储单元，

所述运算单元用于接收控制器单元发出的微指令并进行算数逻辑运算；

所述数据依赖关系判断单元用于对神经元缓存单元进行读写操作，保证指令之间所用的数据不存在读写一致性冲突；

所述神经元缓存单元用于缓存输入神经元数据和输出神经元数据。

所述权值缓存单元用于缓存该从运算模块在计算过程中需要的权值数据；

所述权值梯度缓存单元用于缓存相应从运算模块在更新权值过程中需要的权值梯度数据。

4、如权利要求 1 所述的用于执行循环神经网络和 LSTM 运算的装置，

在循环神经网络和 LSTM 运算的正向过程进行以下计算过程：

以下公式中表达参数的含义：

w_{ij}表示从单位 i 到单位 j 的连接权值，a_j^t表示 t 时刻 j 单位输入门的值，b_j^t表示t 时刻 j 单位输入门的激活值，下标l，$\emptyset$，ω，c 分别表示输入门，输出门，忘记门和第 c 个 cell。w_{cl}，$w_{c\emptyset}$，$w_{c\omega}$分别表示当前第 c 个 cell 到上一时刻的输入门，输出门，忘记门的连接权值，s_c^t表达第 c 个 cell 在 t 时刻的状态，f、g、h 都是激活函数，I 表示输入大小，K 表示输出大小，H 表示隐层大小，h 泛指 cell 和其他时刻的隐层连接，G 表示所有对隐层的输入；

为了简化表达，定义$\delta_j^t = \frac{\partial L}{\partial a_j^t}$。，即残差对于第 j 个 cell 在 t 时刻的偏导数，下面的公式都是在一个时刻下的 block 完成的，其他时刻以此类推；

以下是前向运算的公式表达，计算顺序按照给出公式顺序进行：

输入门：

$$a_l^t = \sum_{i=1}^{I} w_{il} x_i^t + \sum_{h=1}^{H} w_{hl} b_h^t + \sum_{c=1}^{C} w_{cl} s_c^t$$

$$b_l^t = \mathrm{f}(a_l^t)$$

忘记门：

$$a_{\emptyset}^{t} = \sum_{i=1}^{I} w_{i\emptyset} x_i^t + \sum_{h=1}^{H} w_{h\emptyset} b_h^{t-1} + \sum_{c=1}^{C} w_{c\emptyset} s_c^{t-1}$$

$$b_{\emptyset}^{t} = \mathrm{f}(a_{\emptyset}^{t})$$

Cell：

$$a_c^t = \sum_{i=1}^{I} w_{ic} x_i^t + \sum_{h=1}^{H} w_{hc} b_h^{t-1}$$

Cell 状态：

$$s_c^t = b_{\emptyset}^t s_c^{t-1} + b_l^t \mathrm{g}(a_c^t)$$

输出门：

$$a_{\omega}^{t} = \sum_{i=1}^{I} w_{i\omega} x_i^t + \sum_{h=1}^{H} w_{h\omega} b_h^{t-1} + \sum_{c=1}^{C} w_{c\omega} s_c^{t}$$

$$b_{\omega}^{t} = \mathrm{f}(a_{\omega}^{t})$$

Cell 输出：

$$b_c^t = b_{\omega}^t h(s_c^t)$$

。

5、如权利要求 4 所述的用于执行循环神经网络和 LSTM 运算的装置，

在循环神经网络和LSTM运算的单层运算的梯度反向时进行以下计算过程：

反向运算公式表达，计算顺序按照给出公式顺序进行：

定义：

$$\epsilon_c^t = \frac{\partial L}{\partial b_c^t} \qquad \epsilon_s^t = \frac{\partial L}{\partial s_c^t}$$

Cell 输出：

$$\epsilon_c^t = \sum_{k=1}^{K} w_{ck} \delta_k^t + \sum_{g=1}^{G} w_{cg} \delta_g^{t+1}$$

输出门：

$$\delta_{\omega}^{t} = \mathrm{f}'(a_{\omega}^{t}) \sum_{c=1}^{C} h(s_{c}^{t})\, \epsilon_{c}^{t}$$

Cell 状态：

$$\epsilon_{s}^{t} = b_{\omega}^{t} h'(s_{c}^{t}) \epsilon_{c}^{t} + b_{\emptyset}^{t+1} \epsilon_{s}^{t+1} + w_{cl} \delta_{l}^{t+1} + w_{c\emptyset} \delta_{\emptyset}^{t+1} + w_{c\omega} \delta_{\omega}^{t+1}$$

Cell：

$$\delta_{c}^{t} = b_{l}^{t} g'(a_{c}^{t}) \epsilon_{s}^{t}$$

忘记门：

$$\delta_{\emptyset}^{t} = \mathrm{f}'\left(a_{\emptyset}^{t}\right) \sum_{c=1}^{C} s_{c}^{t-1}\, \epsilon_{s}^{t}$$

输入门：

$$\delta_{l}^{t} = \mathrm{f}'(a_{l}^{t}) \sum_{c=1}^{C} g(s_{c}^{t})\, \epsilon_{s}^{t}$$

。

6、一种用于执行循环神经网络和 LSTM 运算的方法，其中，

利用多个从运算模块将输入数据进行乘加得到部分和并保存，直到神经元数据全都输入将结果返回给主运算模块；

利用一个主运算模块在正向过程时对从运算模块返回的和进行插值激活，以及在反向过程时插值得到激活导数并且与梯度相乘。

7、如权利要求 6 所述的用于执行循环神经网络和 LSTM 运算的方法，

在循环神经网络和 LSTM 运算的正向过程进行以下计算过程：

以下公式中表达参数的含义：

w_{ij}表示从单位 i 到单位 j 的连接权值，a_{j}^{t}表示 t 时刻 j 单位输入门的值，b_{j}^{t}表示t 时刻 j 单位输入门的激活值，下标l，$\emptyset$，ω，c 分别表示输入门，输出门，忘记门和第 c 个 cell。w_{cl}，$w_{c\emptyset}$，$w_{c\omega}$分别表示当前第 c 个 cell 到上一时刻的输入门，输出门，忘记门的连接权值，s_{c}^{t}表达第 c 个 cell 在 t 时刻的状态，f、g、h 都是激活函数，I 表示输入大小，

K 表示输出大小，H 表示隐层大小，h 泛指 cell 和其他时刻的隐层连接，G 表示所有对隐层的输入；

为了简化表达，定义$\delta_j^t = \frac{\partial L}{\partial a_j^t}$，即残差对于第 j 个 cell 在 t 时刻的偏导数，下面的公式都是在一个时刻下的 block 完成的，其他时刻以此类推；

以下是前向运算的公式表达，计算顺序按照给出公式顺序进行：

输入门：

$$a_l^t = \sum_{i=1}^{I} w_{il} x_i^t + \sum_{h=1}^{H} w_{hl} b_h^t + \sum_{c=1}^{C} w_{cl} s_c^t$$

$$b_l^t = \mathrm{f}(a_l^t)$$

忘记门：

$$a_\phi^t = \sum_{i=1}^{I} w_{i\phi} x_i^t + \sum_{h=1}^{H} w_{h\phi} b_h^{t-1} + \sum_{c=1}^{C} w_{c\phi} s_c^{t-1}$$

$$b_\phi^t = \mathrm{f}(a_\phi^t)$$

Cell：

$$a_c^t = \sum_{i=1}^{I} w_{ic} x_i^t + \sum_{h=1}^{H} w_{hc} b_h^{t-1}$$

Cell 状态：

$$s_c^t = b_\phi^t s_c^{t-1} + b_l^t \mathrm{g}(a_c^t)$$

输出门：

$$a_\omega^t = \sum_{i=1}^{I} w_{i\omega} x_i^t + \sum_{h=1}^{H} w_{h\omega} b_h^{t-1} + \sum_{c=1}^{C} w_{c\omega} s_c^t$$

$$b_\omega^t = \mathrm{f}(a_\omega^t)$$

Cell 输出：

$$b_c^t = b_\omega^t h(s_c^t)$$

。

8、如权利要求 7 所述的用于执行循环神经网络和 LSTM 运算的装置，

在循环神经网络和LSTM运算的单层运算的梯度反向时进行以下计算过程：

反向运算公式表达，计算顺序按照给出公式顺序进行：

定义：

$$\epsilon_c^t = \frac{\partial L}{\partial b_c^t} \qquad \epsilon_s^t = \frac{\partial L}{\partial s_c^t}$$

Cell 输出：

$$\epsilon_c^t = \sum_{k=1}^{K} w_{ck}\delta_k^t + \sum_{g=1}^{G} w_{cg}\delta_g^{t+1}$$

输出门：

$$\delta_\omega^t = \mathrm{f}'(a_\omega^t)\sum_{c=1}^{C} h(s_c^t)\,\epsilon_c^t$$

Cell 状态：

$$\epsilon_s^t = b_\omega^t h'(s_c^t)\epsilon_c^t + b_\emptyset^{t+1}\epsilon_s^{t+1} + w_{cl}\delta_l^{t+1} + w_{c\emptyset}\delta_\emptyset^{t+1} + w_{c\omega}\delta_\omega^{t+1}$$

Cell：

$$\delta_c^t = b_l^t g'(a_c^t)\epsilon_s^t$$

忘记门：

$$\delta_\emptyset^t = \mathrm{f}'\left(a_\emptyset^t\right)\sum_{c=1}^{C} s_c^{t-1}\,\epsilon_s^t$$

输入门：

$$\delta_l^t = \mathrm{f}'(a_l^t)\sum_{c=1}^{C} g(s_c^t)\,\epsilon_s^t$$

。

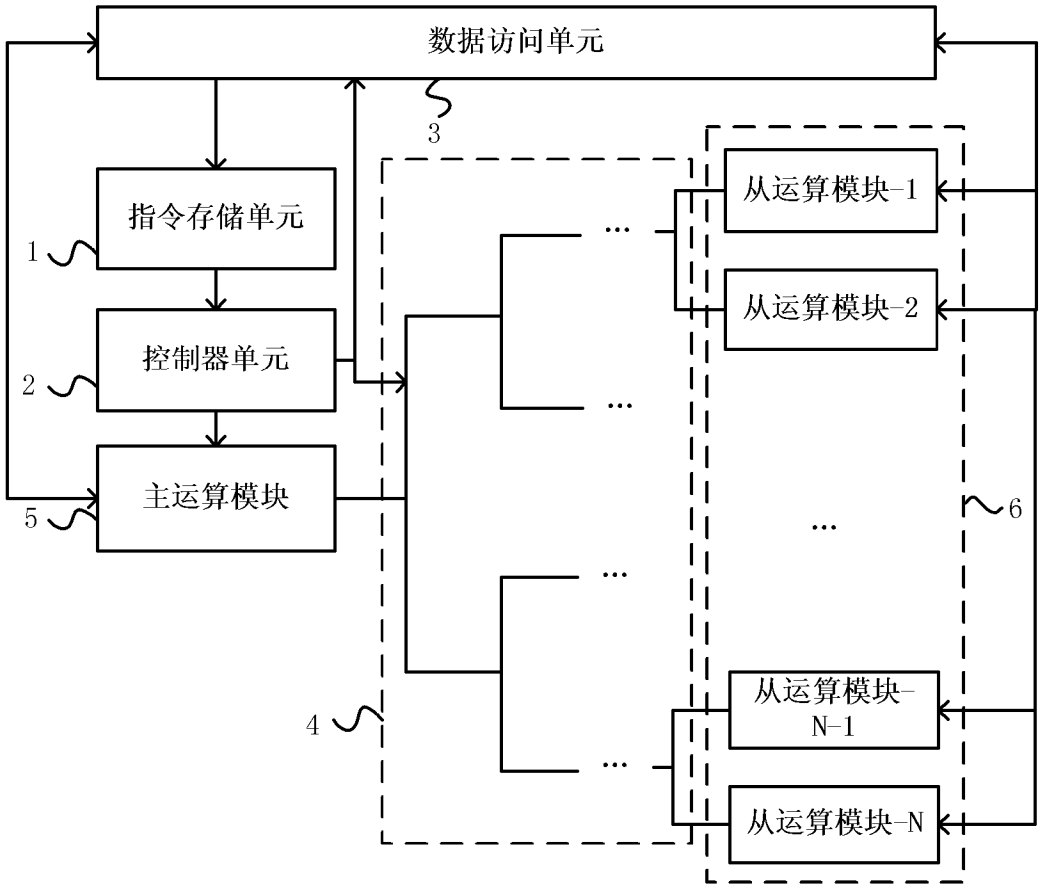
数据访问单元
3
指令存储单元
1
控制器单元
2
主运算模块
5
4
从运算模块-1
从运算模块-2
从运算模块-N-1
从运算模块-N
6

图 1

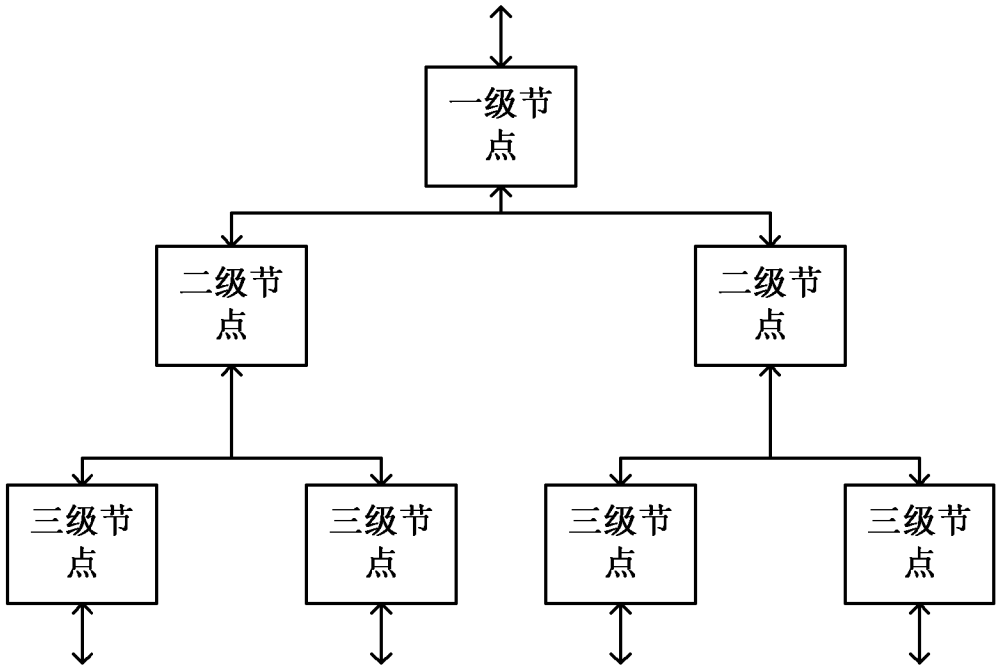
一级节点
二级节点
二级节点
三级节点
三级节点
三级节点
三级节点

图 2

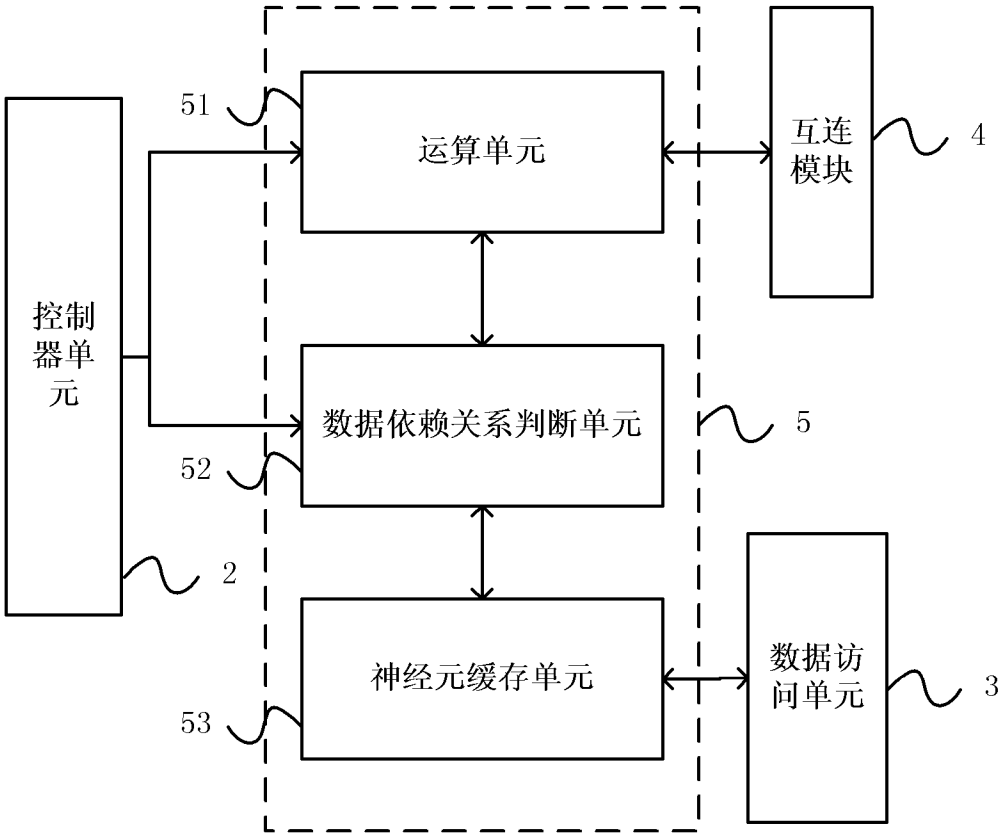
控制器单元
2
51
运算单元
互连模块
4
数据依赖关系判断单元
5
52
神经元缓存单元
53
数据访问单元
3

图 3

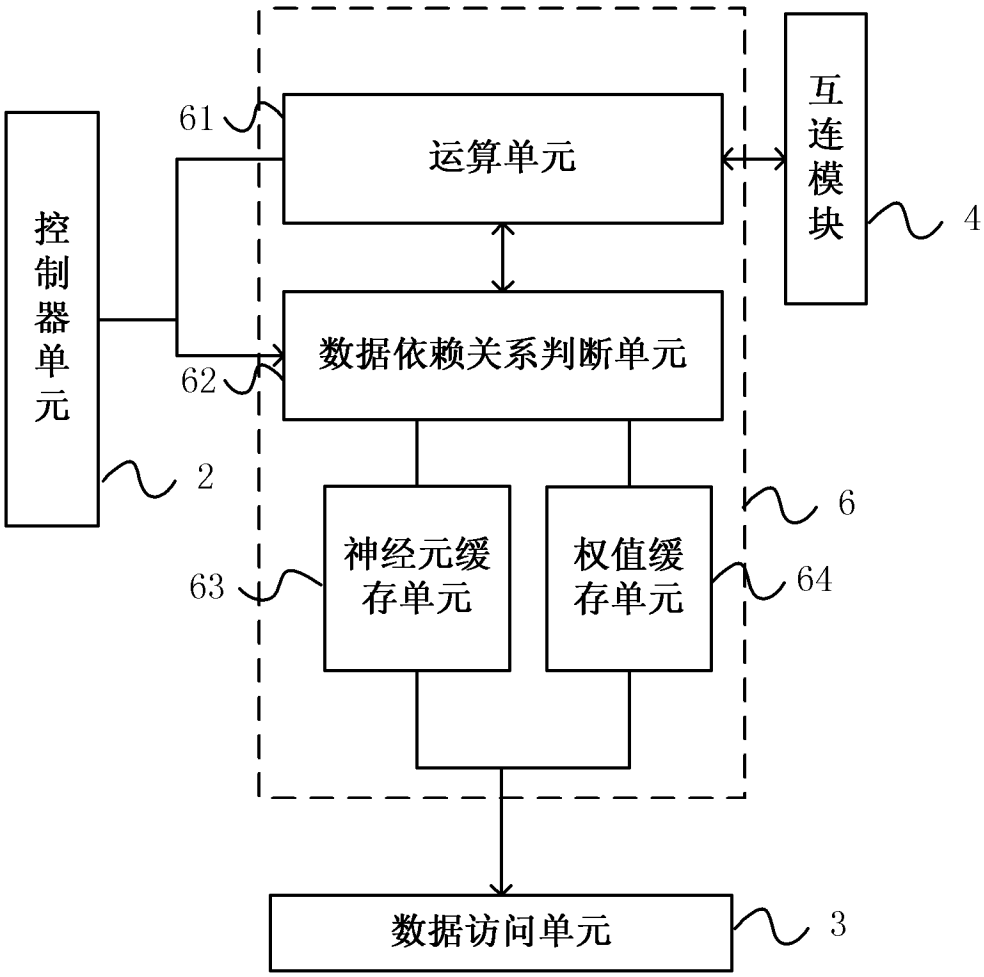
控制器单元
2
61
运算单元
互连模块
4
62
数据依赖关系判断单元
6
63
神经元缓存单元
权值缓存单元
64
数据访问单元
3

图 4

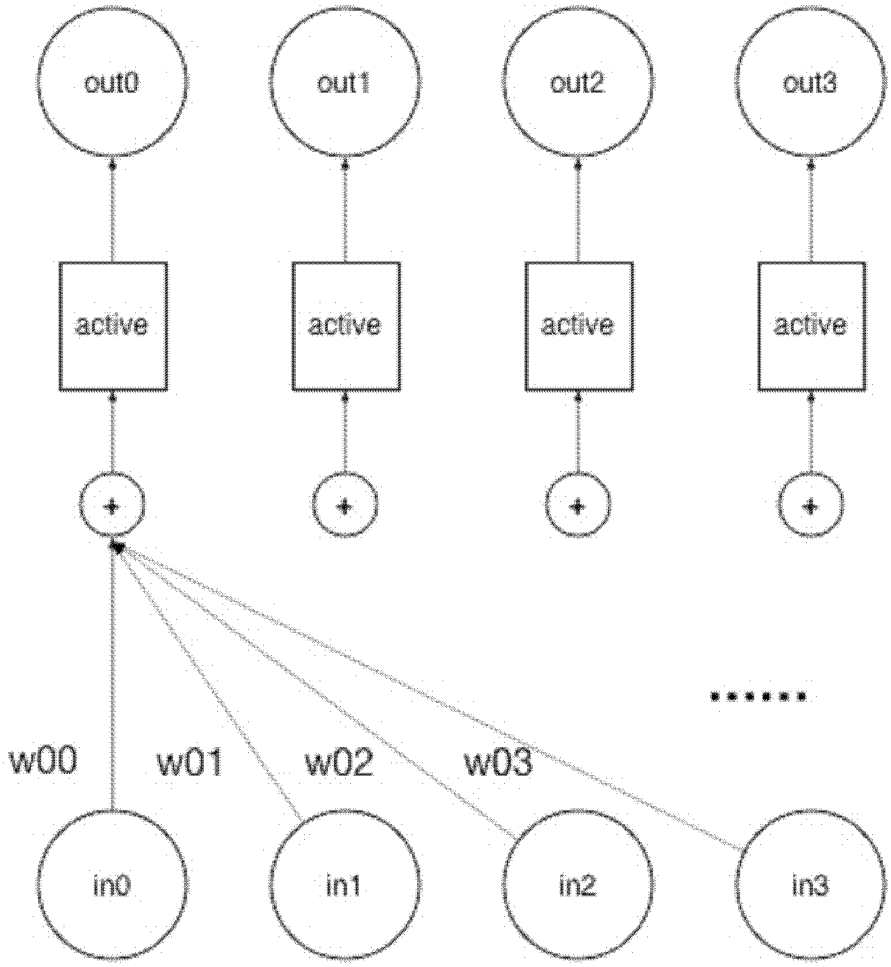
out0
out1
out2
out3
active
active
active
active
+
+
+
+
w00
w01
w02
w03
in0
in1
in2
in3

图 5

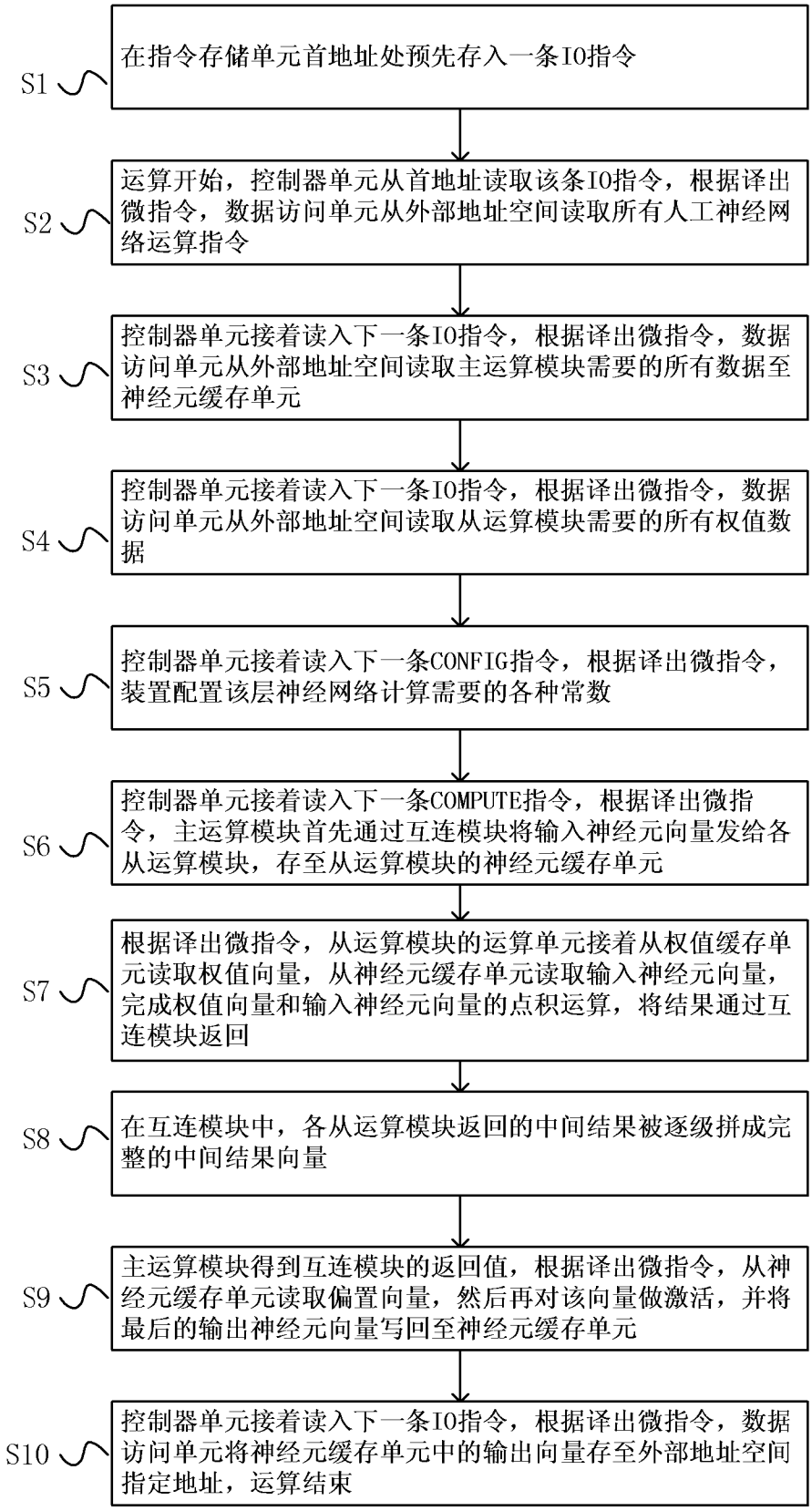
S1
在指令存储单元首地址处预先存入一条IO指令
S2
运算开始，控制器单元从首地址读取该条IO指令，根据译出微指令，数据访问单元从外部地址空间读取所有人工神经网络运算指令
S3
控制器单元接着读入下一条IO指令，根据译出微指令，数据访问单元从外部地址空间读取主运算模块需要的所有数据至神经元缓存单元
S4
控制器单元接着读入下一条IO指令，根据译出微指令，数据访问单元从外部地址空间读取从运算模块需要的所有权值数据
S5
控制器单元接着读入下一条CONFIG指令，根据译出微指令，装置配置该层神经网络计算需要的各种常数
S6
控制器单元接着读入下一条COMPUTE指令，根据译出微指令，主运算模块首先通过互连模块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，将结果通过互连模块返回
S8
在互连模块中，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
主运算模块得到互连模块的返回值，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，数据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 6

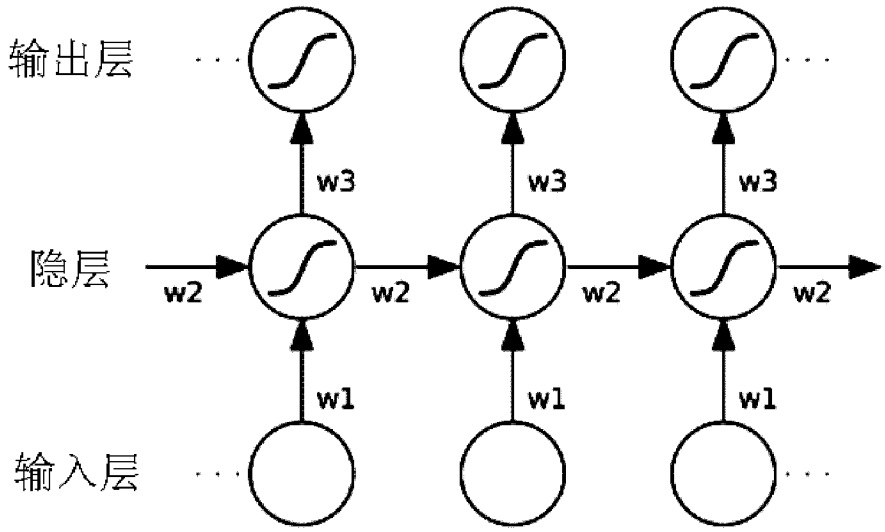

图 7

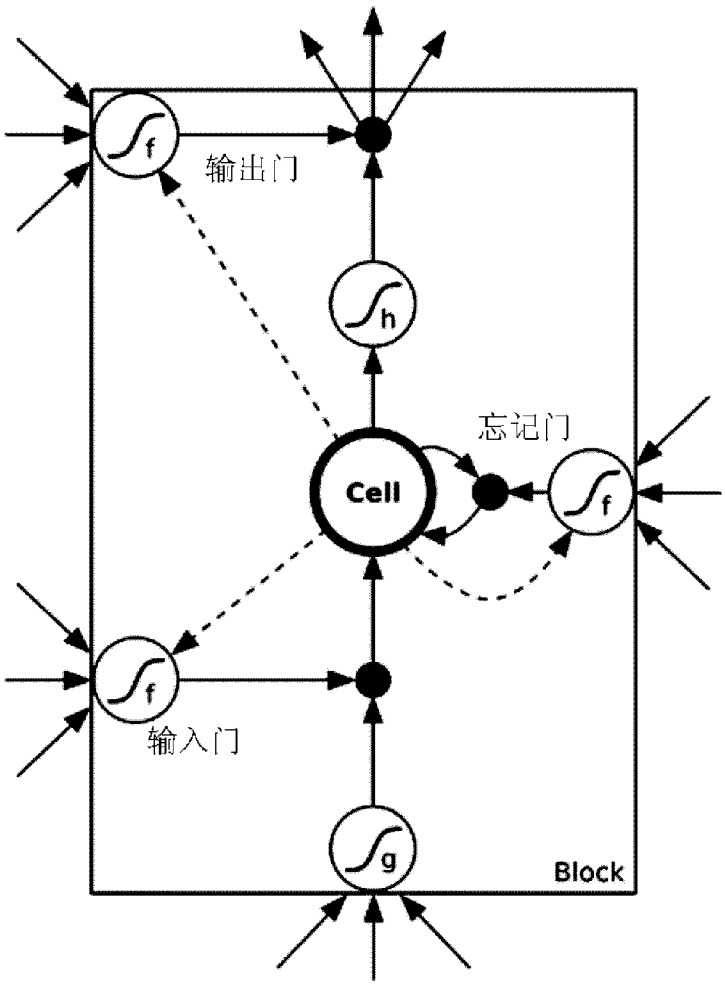

图 8

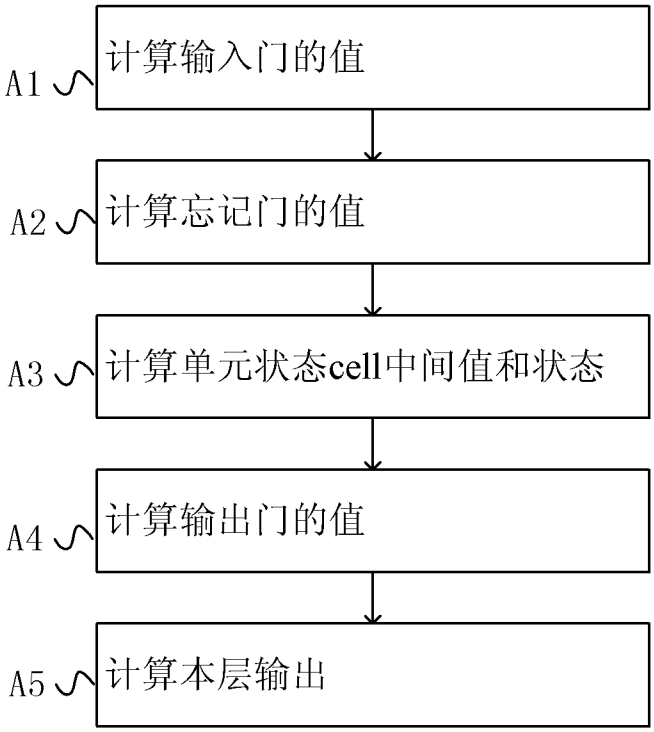
A1
计算输入门的值
A2
计算忘记门的值
A3
计算单元状态cell中间值和状态
A4
计算输出门的值
A5
计算本层输出

图 9

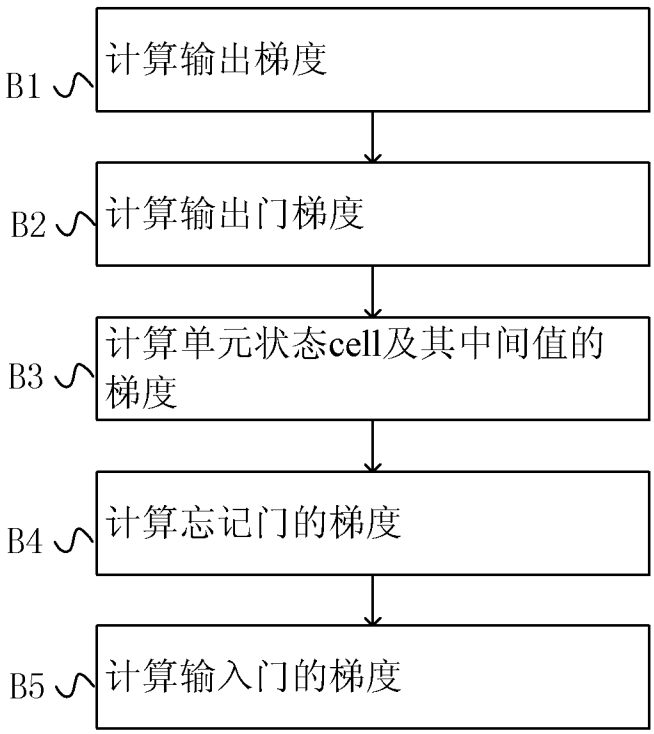
B1
计算输出梯度
B2
计算输出门梯度
B3
计算单元状态cell及其中间值的梯度
B4
计算忘记门的梯度
B5
计算输入门的梯度

图 10

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2016/080744

A. CLASSIFICATION OF SUBJECT MATTER

G06N 3/063 (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, G10L

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)

CNABS, CNTXT, DWPI, CNKI, BAIDU SCHOLAR, GOOGLE SCHOLAR: GUO, Qi; CHEN, Yunji; CAMBRICON; LOONGSON TECHNOLOGY; INSTITUTE OF COMPUTING TECHNOLOGY; multiple, service, calculate, handle, cooperation, main, slave, Recurrent Neural Network, long short term memory, RNN, LSTM

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 102947818 A (THE REGENTS OF THE UNIVERSITY OF CALIFORNIA), 27 February 2013 (27.02.2013), description, paragraphs [0036]-[0048] and [0056]-[0071], and figures 2 and 5-6	1, 4-8
A	CN 102947818 A (THE REGENTS OF THE UNIVERSITY OF CALIFORNIA), 27 February 2013 (27.02.2013), description, paragraphs [0036]-[0048] and [0056]-[0071], and figures 2 and 5-6	2-3
A	JP 07311754 A (RICOH KK.), 28 November 1995 (28.11.1995), the whole document	1-8
A	CN 105488565 A (INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES), 13 April 2016 (13.04.2016), the whole document	1-8
A	CN 103970512 A (LOONGSON TECHNOLOGY CORPORATION LIMITED), 06 August 2014 (06.08.2014), the whole document	1-8
A	CN 104145281 A (AN, Bingyi), 12 November 2014 (12.11.2014), the whole document	1-8
A	CN 105512723 A (NANJING AIXI INFORMATION TECHNOLOGY CO., LTD.), 20 April 2016 (20.04.2016), the whole document	1-8

☒ 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
22 January 2017 (22.01.2017)	**06 February 2017 (06.02.2017)**
Name and mailing address of the ISA/CN: 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 **YOU, Xiaomei** Telephone No.: (86-10) **62089539**

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2016/080744

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 104538028 A (TSINGHUA UNIVERSITY), 22 April 2015 (22.04.2015), the whole document	1-8
A	CHEN, Yunji et al. "Artificial Neural Network Processor." SCIENCE CHINA: CHINESE BULLETIN OF LIFE SCIENCE, vol. 46, no. 2, 27 February 2016 (27.02.2016), ISSN: 1674-7232, pages 223-224.	1-8

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2016/080744

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 102947818 A	27 February 2013	US 8655815 B2	18 February 2014
		EP 2572293 A1	27 March 2013
		JP 2013529342 A	18 July 2013
		WO 2011146147 A1	24 November 2011
		US 2011289034 A1	24 November 2011
		CN 102947818 B	22 July 2015
		KR 20130111956 A	11 October 2013
		CA 2799167 A1	24 November 2011
		EP 2572293 A4	04 December 2013
		US 2014172763 A1	19 June 2014
		HK 1180077 A0	11 October 2013
		IN 201210441 P4	23 May 2014
JP 07311754 A	28 November 1995	None	
CN 105488565 A	13 April 2016	None	
CN 103970512 A	06 August 2014	CN 103970512 B	14 September 2016
CN 104145281 A	12 November 2014	US 2014344203 A1	20 November 2014
		WO 2013115431 A1	08 August 2013
		KR 20130090147 A	13 August 2013
CN 105512723 A	20 April 2016	None	
CN 104538028 A	22 April 2015	WO 2016101688 A1	30 June 2016

国际检索报告

国际申请号
PCT/CN2016/080744

A. 主题的分类

G06N 3/063(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，G10L

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））

CNABS,CNTXT,DWPI,CNKI,百度学术，谷歌学术:循环神经网络，长短时记忆，递归神经网络，回归神经网络，郭崎，陈云霁，中科寒武纪，龙芯科技，计算技术研究所，主，从，多，服务，计算，处理，协同，main,slave, Recurrent Neural Network,long short term memory,RNN,LSTM

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
X	CN 102947818 A (加利福尼亚大学董事会) 2013年 2月 27日 (2013 - 02 - 27) 说明书第[0036]-[0048]，[0056]-[0071]段，图2，图5-6	1，4-8
A	CN 102947818 A (加利福尼亚大学董事会) 2013年 2月 27日 (2013 - 02 - 27) 说明书第[0036]-[0048]，[0056]-[0071]段，图2，图5-6	2-3
A	JP 07311754 A (RICOH KK.) 1995年 11月 28日 (1995 - 11 - 28) 全文	1-8
A	CN 105488565 A (中国科学院计算技术研究所) 2016年 4月 13日 (2016 - 04 - 13) 全文	1-8
A	CN 103970512 A (龙芯中科技术有限公司) 2014年 8月 6日 (2014 - 08 - 06) 全文	1-8
A	CN 104145281 A (安秉益) 2014年 11月 12日 (2014 - 11 - 12) 全文	1-8
A	CN 105512723 A (南京艾溪信息科技有限公司) 2016年 4月 20日 (2016 - 04 - 20) 全文	1-8

☑ 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
2017年 1月 22日	2017年 2月 6日
ISA/CN的名称和邮寄地址 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 传真号 (86-10)62019451	受权官员 游晓梅 电话号码 (86-10)62089539

表 PCT/ISA/210 (第2页) (2009年7月)

国际检索报告

国际申请号
PCT/CN2016/080744

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
A	CN 104538028 A (清华大学) 2015年 4月 22日 (2015 - 04 - 22) 全文	1-8
A	陈云霁 等. "人工神经网络处理器." 中国科学:生命科学., 第46卷，第2期，2016年 2月 27日 (2016 - 02 - 27)， ISSN：1674-7232, 第223-224页.	1-8

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2016/080744

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	102947818	A	2013年 2月 27日	US	8655815	B2	2014年 2月 18日
				EP	2572293	A1	2013年 3月 27日
				JP	2013529342	A	2013年 7月 18日
				WO	2011146147	A1	2011年 11月 24日
				US	2011289034	A1	2011年 11月 24日
				CN	102947818	B	2015年 7月 22日
				KR	20130111956	A	2013年 10月 11日
				CA	2799167	A1	2011年 11月 24日
				EP	2572293	A4	2013年 12月 4日
				US	2014172763	A1	2014年 6月 19日
				HK	1180077	A0	2013年 10月 11日
				IN	201210441	P4	2014年 5月 23日
JP	07311754	A	1995年 11月 28日		无		
CN	105488565	A	2016年 4月 13日		无		
CN	103970512	A	2014年 8月 6日	CN	103970512	B	2016年 9月 14日
CN	104145281	A	2014年 11月 12日	US	2014344203	A1	2014年 11月 20日
				WO	2013115431	A1	2013年 8月 8日
				KR	20130090147	A	2013年 8月 13日
CN	105512723	A	2016年 4月 20日		无		
CN	104538028	A	2015年 4月 22日	WO	2016101688	A1	2016年 6月 30日