(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 2017/185386 A1

(51) 国际专利分类号:
G06F 9/30 (2006.01)

(21) 国际申请号: PCT/CN2016/080967

(22) 国际申请日: 2016 年 5 月 4 日 (04.05.2016)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201610282534.7 2016年4月29日 (29.04.2016) CN

(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: 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)。 韩栋 (HAN, Dong); 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)。 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)。 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)。

(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,

(54) **Title:** DEVICE AND METHOD FOR PERFORMING FORWARD OPERATION OF CONVOLUTIONAL NEURAL NETWORK

(54) 发明名称: 一种用于执行卷积神经网络正向运算的装置和方法

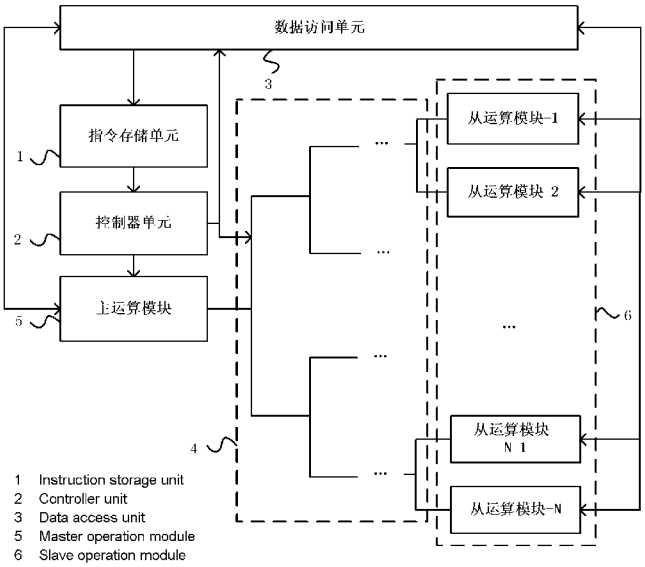

图 3

(57) **Abstract:** A device for performing a convolutional neural network, comprising an instruction storage unit (1), a controller unit (2), a data access unit (3), an interconnection module (4), a master operation module (5), and a plurality of slave operation modules (6). The device can implement a forward operation of one or more convolutional layers of an artificial neural network. For each layer, first, data is selected from an input neuron vector according to a convolution window, then convolution operation is performed with a convolution kernel to calculate an intermediate result of the layer, and then the intermediate result is biased and activated to obtain output data. The output data is used as input data of a next layer.

WO 2017/185386 A1

[见续页]

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, 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：一种执行卷积神经网络的装置，其中装置部分包括了指令存储单元（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）。使用该装置可以实现一层或多层人工神经网络卷积层的正向运算。对于每一层来说，首先对输入神经元向量依据卷积窗口进行数据选择，然后与卷积核进行卷积运算，计算出本层的中间结果，然后对该中间结果加偏置并激活得到输出数据。将输出数据作为下一层的输入数据。

一种用于执行卷积神经网络正向运算的装置和方法

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卷积神经网络正向运算的装置和方法。

背景技术

卷积神经网络是近年来广泛应用于模式识别、图像处理等领域的一种高效识别算法，它具有结构简单、训练参数少和适应性强、平移、旋转、缩放等特点。由于 CNN/DNN 的特征检测层通过训练数据进行学习，所以在使用 CNN/DNN 时，避免了显示的特征抽取，而隐式地从训练数据中进行学习；再者由于同一特征映射面上的神经元权值相同，所以网络可以并行学习，这也是卷积网络相对于神经元彼此相连网络的一大优势。

在已有的计算机领域应用中，与卷积运算相关的应用十分普遍。本发明专注于卷积神经网络，目前可以执行此种运算的主流装置如下：

在现有技术中，一种进行卷积神经网络运算的已知方案是使用通用处理器，该方法通过通用寄存器堆和通用功能部件来执行通用指令，从而执行卷积神经网络运算。然而，该方法的缺点之一是单个通用处理器多用于标量计算，在进行卷积神经网络运算时运算性能较低。而使用多个通用处理器并行执行时，通用处理器之间的相互通讯又有可能成为性能瓶颈。

在另一种现有技术中，使用图形处理器（GPU）来进行向量计算，其中，通过使用通用寄存器堆和通用流处理单元执行通用 SIMD 指令来进行卷积神经网络运算。然而，上述方案中，GPU 片上缓存太小，在进行大规模卷积神经网络运算时需要不断进行片外数据搬运，片外带宽成为了主要性能瓶颈。

发明内容

（一）要解决的技术问题

本发明的目的在于，提供一种支持卷积神经网络的装置，解决现有技术中存在的受限于片间通讯、片上缓存不够等问题。

（二）技术方案

本发明的一个方面提供了一种用于执行卷积神经网络正向运算的装置，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：

指令存储单元通过数据访问单元读入指令并存储读入的指令；

控制器单元从指令存储单元中读取指令，将指令译成控制其他模块行为的控制信号，所述其他模块包括数据访问单元、主运算模块和所述多个从运算模块；

数据访问单元执行外部地址空间与所述装置之间的数据或指令读写操作；

从运算模块用于实现卷积神经网络算法中的输入数据和卷积核的卷积运算；

互连模块用于所述主运算模块和所述从运算模块之间的数据传输，在神经网络全连接层正向运算开始之前，主运算模块通过互连模块将输入数据输送到每一个从运算模块，在从运算模块的计算过程结束后，互连模块逐级将各从运算模块的输出标量拼成中间向量，输送回主运算模块；

主运算模块将所有输入数据的中间向量拼接成中间结果，并对所述中间结果执行后续运算。

本发明的另一方面提供了一种使用上述装置执行单层人工神经网络卷积层正向运算的方法。

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卷积层正向运算的方法。

（三）有益效果

本发明提供的卷积神经网络运算装置及配套指令，将参与计算的输入数据和卷积核暂存在高速暂存存储器上（Scratchpad Memory）。在仅发送同一条指令的情况下，卷积神经网络运算单元中可以更加灵活有效地支持不同宽度的数据，并可以解决数据存储中的相关性问题，从而提升了包含大量卷积神经网络计算任务的执行性能，本发明采用的指令具有精简的格式，使得指令集使用方便、支持的向量长度灵活。

本发明可以应用于以下（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 是卷积神经网络算法的示意流程图。

图 2 是根据本发明实施例的支持卷积神经网络正向运算的装置的指令示意图。

图 3 示出了根据本发明实施例的用于执行卷积神经网络正向运算的装置的整体结构的示例框图。

图 4 示意性示出了根据本发明实施例的用于执行卷积神经网络正向运算的装置中 H 树模块（互连模块一种实现方式）的结构。

图 5 示出了根据本发明实施例的用于执行卷积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 6 示出了根据本发明实施例的用于执行卷积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 7 示出了根据本发明实施例的单层卷积神经网络正向运算过程的示例框图。

具体实施方式

本发明提供一种卷积神经网络计算装置及配套指令集，包括存储单元、寄存器单元和卷积神经网络运算单元，存储单元中存储有输入输出数据和卷积核，寄存器单元中存储有输入输出数据和卷积核存储的地址，卷积神经网络运算单元根据卷积神经网络运算指令在寄存器单元中获取数据地址，然后，根据该数据地址在存储单元中获取相应的输入数据和卷积核，接着，根据获取的输入数据和卷积核进行卷积神经网络运算，得到卷积神经网络运算结果。本发明将参与计算的输入数据和卷积核暂存在外部存储空间（例如，高速暂存存储器）上，使得卷积神经网络运算过程中可以更加灵活有效地支持不同宽度的数据，提升包含大量卷积神经网络计算任务的执行性能。

图 1 为卷积神经网络算法示意图，如图 1 所示，卷积神经网络包括输出数据及激活函数，输入数据层，和卷积核。

首先，其每一次的计算过程，首先需要依据卷积窗口，选取出输入数据层中对应的输入数据 x_i，然后将输入数据和卷积核进行加和运算。其输出数据的计算过程为 $s = s(\sum wx_i + b)$，即将卷积核 w 乘以输入数据 x_i，进行求和，然后加上偏置 b 后做激活运算 $s(h)$，得到最终的输出数据 s。其中，卷积核和输入数据的乘法是向量乘法。

卷积窗口依据卷积核在 X 轴上的大小k_x和在 Y 轴上的大小k_y，在 X 轴尺寸为 W 和 Y 轴尺寸为 H 的输入数据上，从最开始选取出和卷积核同样大小的输入数据，然后

依据卷积窗口的平移位矢S_x和S_y，以此先作水平平移，然后再做垂直平移，将全部输入数据做遍历。

图 2 是根据本发明实施例的指令集的格式示意图，如图 2 所示，卷积神经网络运算指令包括至少 1 个操作码和至少 1 个操作域，其中，操作码用于指示该卷积神经网络运算指令的功能，卷积神经网络运算单元通过识别该操作码可进行卷积神经网络运算，操作域用于指示该卷积神经网络运算指令的数据信息，其中，数据信息可以是立即数或寄存器号，包括输入数据的起始地址和数据长度，卷积核的起始地址和数据长度，以及激活函数的类型。

指令集包含有不同功能的卷积神经网络 COMPUTE 指令以及 CONFIG 指令、IO 指令、NOP 指令、JUMP 指令和 MOVE 指令。在一种实施例中，COMPUTE 指令包括：

卷积神经网络 sigmoid 指令，根据该指令，装置分别从高速暂存存储器的指定地址取出指定大小的输入数据和卷积核，在卷积运算部件中做卷积操作，然后将输出结果做 sigmoid 激活；

卷积神经网络 TanH 指令，根据该指令，装置分别从高速暂存存储器的指定地址取出指定大小的输入数据和卷积核，在卷积运算部件中做卷积操作，然后将输出结果做 TanH 激活；

卷积神经网络 ReLU 指令，根据该指令，装置分别从高速暂存存储器的指定地址取出指定大小的输入数据和卷积核，在卷积运算部件中做卷积操作，然后将输出结果做 ReLU 激活；以及

卷积神经网络 group 指令，根据该指令，装置分别从高速暂存存储器的指定地址取出指定大小的输入数据和卷积核，划分 group 之后，在卷积运算部件中做卷积操作，然后将输出结果做激活。

COMPUTE 指令也可以包括其他的运算指令，进行非线性激活和线性激活操作。

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。

IO 指令实现从外部存储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指令完毕。NOP 指令本身不包含任何操作；

JUMP 指令负责控制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。

为使本发明的目的、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，对本发明进一步详细说明。

图 3 是本发明实施例提供的卷积神经网络正向运算装置的结构示意图。如图 3 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（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。

指令存储单元 1 通过数据访问单元 3 读入指令并存储读入的指令。

控制器单元 2 从指令存储单元 1 中读取指令，将指令译成控制其他模块行为的控制信号并发送给其他模块如数据访问单元 3、主运算模块 5 和从运算模块 6 等。

数据访问单元 3 能够访问外部地址空间，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。

互连模块 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 4 示意性示出了互连模块 4 的一种实施方式：H 树模块。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，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，在卷积神经网络开始计算阶段，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，当从运算模块的计算过程完成后，每个从运算模块输出的神经元的值会在互连模块中逐级拼成一个完整的由神经元组成的向量。举例说明，假设装置中共有 N 个从运算模块，则输入数据 x_i 被发送到 N 个从运算模块，每个从运算模块将输入数据 x_i 与该从运算模块相应的卷积核做卷积运算，得到一标量数据，各从运算模块的标量数据被互连模块 4 合并成一个含有 N 个元素的中间向量。假设卷积窗口总共遍历得到 A*B 个（X 方向为 A 个，Y 方向为 B 个，X、Y 为三维正交坐标系的坐标轴）输入数据 x_i，则对 A*B 个 x_i 执行上述卷积操作，得到的所有向量在主运算模块中合并得到 A*B*N 的三维中间结果。

图 5 示出了根据本发明实施例的用于执行卷积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 5 所示，主运算模块 5 包括第一运算单元 51、第一数据依赖关系判定单元 52 和第一存储单元 53。

其中，第一运算单元 51 包括向量加法单元 511 以及激活单元 512。第一运算单元 51 接收来自控制器单元的控制信号，完成主运算模块 5 的各种运算功能，向量加法单元 511 用于实现卷积神经网络正向计算中的加偏置操作，该部件将偏置数据与所述中间结果对位相加得到偏置结果，激活运算单元 512 对偏置结果执行激活函数操作。所述偏置数据可以是从外部地址空间读入的，也可以是存储在本地的。

第一数据依赖关系判定单元 52 是第一运算单元 51 读写第一存储单元 53 的端口，保证第一存储单元 53 中数据的读写一致性。同时，第一数据依赖关系判定单元 52 也负责将从第一存储单元 53 读取的数据通过互连模块 4 发送给从运算模块，而从运算模块 6 的输出数据通过互连模块 4 直接发送给第一运算单元 51。控制器单元 2 输出的指令发送给计算单元 51 和第一数据依赖关系判定单元 52，来控制其行为。

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数据和输出数据。

图 6 示出了根据本发明实施例的用于执行卷积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 包括第二运算单元 61、数据依赖关系判定单元 62、第二存储单元 63 和第三存储单元 64。

第二运算单元 61 接收控制器单元 2 发出的控制信号并进行卷积运算。第二运算单元包括向量乘单元 611 和累加单元 612，分别负责卷积运算中的向量乘运算和累加运算。

第二数据依赖关系判定单元 62 负责计算过程中对第二存储单元 63 的读写操作。第二数据依赖关系判定单元 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 的控制信号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。

第二存储单元 63 缓存该从运算模块 6 的输入数据和输出标量数据。

第三存储单元 64 缓存该从运算模块 6 在计算过程中需要的卷积核数据。

图 7 是本发明实施例提供的卷积神经网络运算装置执行卷积神经网络的流程图，如图 7 所示，执行卷积神经网络指令的过程包括：

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，根据译出的控制信号，数据访问单元 3 从外部地址空间读取相应的所有卷积神经网络运算指令，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，根据译出的控制信号，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，包括输入数据、用于作快速的激活函数运算的插值表、用于配置运算器件参数的常数表、偏置数据等）至主运算模块 5 的第一存储单元 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，根据译出的控制信号，数据访问单元 3 从外部地址空间读取从运算模块 6 需要的卷积核数据。

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，根据译出的控制信号，装置配置该层神经网络计算需要的各种常数。例如，第一运算单元 51、第二运算单元 61 根据控制信号里的参数配置单元内部寄存器的值，所述参数包括例如激活函数需要的数据。

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，根据译出的控制信号，主运算模块 5 首先通过互连模块 4 将卷积窗口内的输入数据发给各从运算模块 6，保存至从运算模块 6 的第二存储单元 63，之后，再依据指令移动卷积窗口。

在步骤 S7，根据 COMPUTE 指令译出的控制信号，从运算模块 6 的运算单元 61 从第三存储单元 64 读取卷积核，从第二存储单元 63 读取输入数据，完成输入数据和卷积核的卷积运算，将中间结果通过互连模块 4 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向量。

在步骤 S9，主运算模块 5 得到互连模块 4 返回的中间向量，卷积窗口遍历所有输入数据，主运算模块将所有返回向量拼接成中间结果，根据 COMPUTE 指令译出的控制信号，从第一存储单元 53 读取偏置数据，与中间结果通过向量加单元 511 相加得到偏置结果，然后激活单元 512 对偏置结果做激活，并将最后的输出数据写回至第一存储单元 53 中。

在步骤 S10，控制器单元 2 接着从指令存储单元读入下一条 IO 指令，根据译出的控制信号，数据访问单元 3 将第一存储单元 53 中的输出数据存至外部地址空间指定地址，运算结束。

对于多层神经网络卷积层，其实现过程与单层神经网络卷积层类似，当上一层卷积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数据地址作为本层的输入数据地址。同样地，指令中的卷积核和偏置数据地址也会变更至本层对应的地址。

通过采用用于执行卷积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。有效提高了对多层卷积神经网络正向运算的支持。

通过采用针对多层卷积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卷积核数据的重用性，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卷积神经网络正向运算性能瓶颈的问题。

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。

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、一种用于执行卷积神经网络正向运算的装置，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：

指令存储单元通过数据访问单元读入指令并存储读入的指令；

控制器单元从指令存储单元中读取指令，将指令译成控制其他模块行为的控制信号，所述其他模块包括数据访问单元、主运算模块和所述多个从运算模块；

数据访问单元执行外部地址空间与所述装置之间的数据或指令读写操作；

从运算模块用于实现卷积神经网络算法中的输入数据和卷积核的卷积运算；

互连模块用于所述主运算模块和所述从运算模块之间的数据传输，在神经网络全连接层正向运算开始之前，主运算模块通过互连模块将输入数据输送到每一个从运算模块，在从运算模块的计算过程结束后，互连模块逐级将各从运算模块的输出标量拼成中间向量，输送回主运算模块；

主运算模块将所有输入数据的中间向量拼接成中间结果，并对所述中间结果执行后续运算。

2、根据权利要求 1 所述的装置，其中主运算模块将中间结果与偏置数据相加，然后执行激活操作。

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输入数据和各自的卷积核，并行地计算出各自的输出标量。

4、根据权利要求 1 所述的装置，其中主运算模块使用的激活函数 active 是非线性函数 sigmoid，tanh，relu，softmax 中的任一个或线性函数。

5、根据权利要求 1 所述的装置，其中，互连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，互连模块为以下任一种结构：树状结构、环状结构、网格状结构、分级互连、总线结构。

6、根据权利要求 1 所述的装置，其中，主运算模块包括第一存储单元、第一运算单元、第一数据依赖关系判定单元和第一存储单元，其中：

第一存储单元用于缓存主运算模块在计算过程中用到的输入数据和输出数据；

第一运算单元完成主运算模块的各种运算功能；

数据依赖关系判定单元是第一运算单元读写第一存储单元的端口，保证对第一存储单元的数据读写不存在一致性冲突，并且负责从第一存储单元读取输入的神经元向量，并通过互连模块发送给从运算模块；以及

来自互连模块的中间向量被发送到第一运算单元。

7、根据权利要求 1 所述的装置，其中，每个从运算模块包括第二运算单元、第二数据依赖关系判定单元、第二存储单元和第三存储单元，其中：

第二运算单元接收控制器单元发出的控制信号并进行算数逻辑运算；

第二数据依赖关系判定单元负责计算过程中对第二存储单元和第三存储单元的读写操作，保证对第二存储单元和第三存储单元的读写不存在一致性冲突；

第二存储单元缓存输入数据以及该从运算模块计算得到的输出标量；以及

第三存储单元缓存该从运算模块在计算过程中需要的卷积核。

8、根据权利要求 6 或 7 所述的装置，其中，第一和第二数据依赖关系判定单元通过以下方式保证读写不存在一致性冲突：判断尚未执行的控制信号与正在执行过程中的控制信号的数据之间是否存在依赖关系，如果不存在，允许该条控制信号立即发射，否则需要等到该条控制信号所依赖的所有控制信号全部执行完成后该条控制信号才允许被发射。

9、根据权利要求 1 所述的装置，其中，数据访问单元从外部地址空间读入以下中的至少一个：输入数据、偏置数据、或卷积核。

10、一种用于执行单层卷积神经网络正向运算的方法，包括：

在步骤 S1，在指令存储单元的首地址处预先存入一条 IO 指令；

在步骤 S2，运算开始，控制器单元从指令存储单元的首地址读取该条 IO 指令，根据译出的控制信号，数据访问单元从外部地址空间读取相应的所有卷积神经网络运算指令，并将其缓存在指令存储单元中；

在步骤 S3，控制器单元接着从指令存储单元读入下一条 IO 指令，根据译出的控制信号，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主运算模块的第一存储单元；

在步骤 S4，控制器单元接着从指令存储单元读入下一条 IO 指令，根据译出的控制信号，数据访问单元从外部地址空间读取从运算模块需要的卷积核数据；

在步骤 S5，控制器单元接着从指令存储单元读入下一条 CONFIG 指令，根据译出的控制信号，装置配置该层神经网络计算需要的各种常数；

在步骤 S6，控制器单元接着从指令存储单元读入下一条 COMPUTE 指令，根据译出的控制信号，主运算模块首先通过互连模块将卷积窗口内的输入数据发给各从运算模块，保存至从运算模块的第二存储单元，之后，在依据指令移动卷积窗口；

在步骤S7，根据COMPUTE指令译出的控制信号，从运算模块的运算单元从第三存储单元读取卷积核，从第二存储单元读取输入数据，完成输入数据和卷积核的卷积运算，将得到的输出标量通过互连模块返回；

在步骤S8，在互连模块中，各从运算模块返回的输出标量被逐级拼成完整的中间向量；

在步骤S9，主运算模块得到互连模块返回的中间向量，卷积窗口遍历所有输入数据，主运算模块将所有返回向量拼接成中间结果，根据COMPUTE指令译出的控制信号，从第一存储单元读取偏置数据，与中间结果通过向量加单元相加得到偏置结果，然后激活单元对偏置结果做激活，并将最后的输出数据写回至第一存储单元中；

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，根据译出的控制信号，数据访问单元将第一存储单元中的输出数据存至外部地址空间指定地址，运算结束。

11、一种用于执行多层卷积神经网络正向运算的方法，包括：

对每一层执行根据权利要求10所述的方法，当上一层卷积神经网络执行完毕后，本层的运算指令将主运算单元中存储的上一层的输出数据地址作为本层的输入数据地址，并且指令中的卷积核和偏置数据地址变更至本层对应的地址。

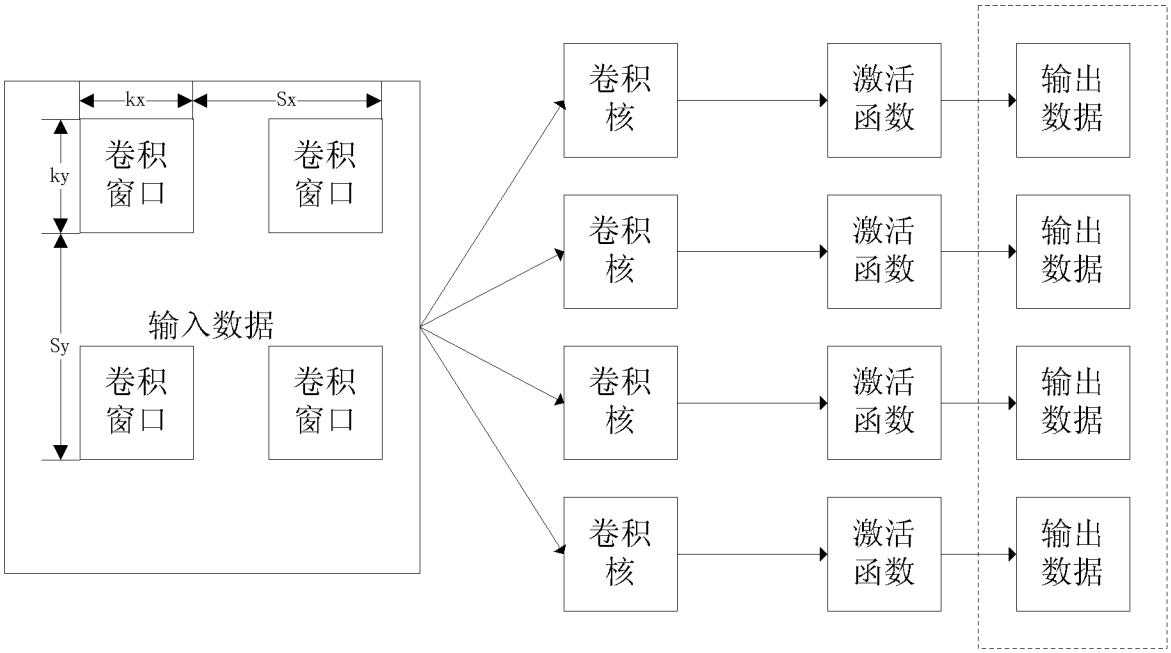

图 1

操作码	寄存器堆 0	寄存器堆 1	寄存器堆 2	寄存器堆 3	寄存器堆 4
COMPUTE	输入数据起始地址	输入数据长度	卷积核起始地址	卷积核长度	激活函数插值表地址
IO	数据外部存储其地址	数据长度	数据内部存储器地址		
NOP					
JUMP	目标地址				
MOVE	输入地址	数据大小	输出地址		

图 2

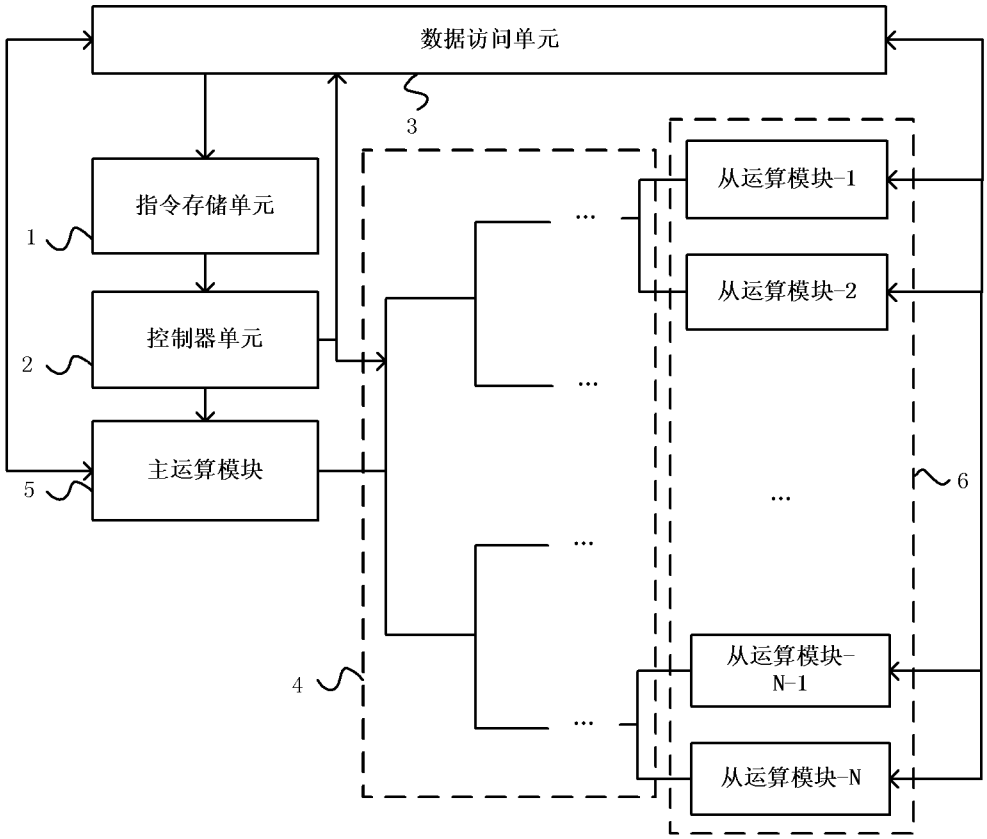
数据访问单元
3
指令存储单元
1
控制器单元
2
主运算模块
5
4
从运算模块-1
从运算模块-2
从运算模块-N-1
从运算模块-N
6

图 3

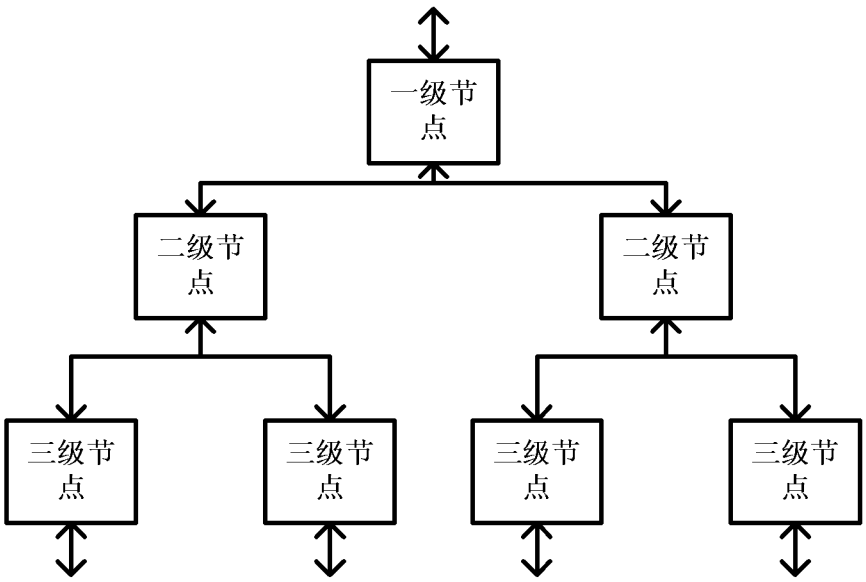
一级节点
二级节点
二级节点
三级节点
三级节点
三级节点
三级节点

图 4

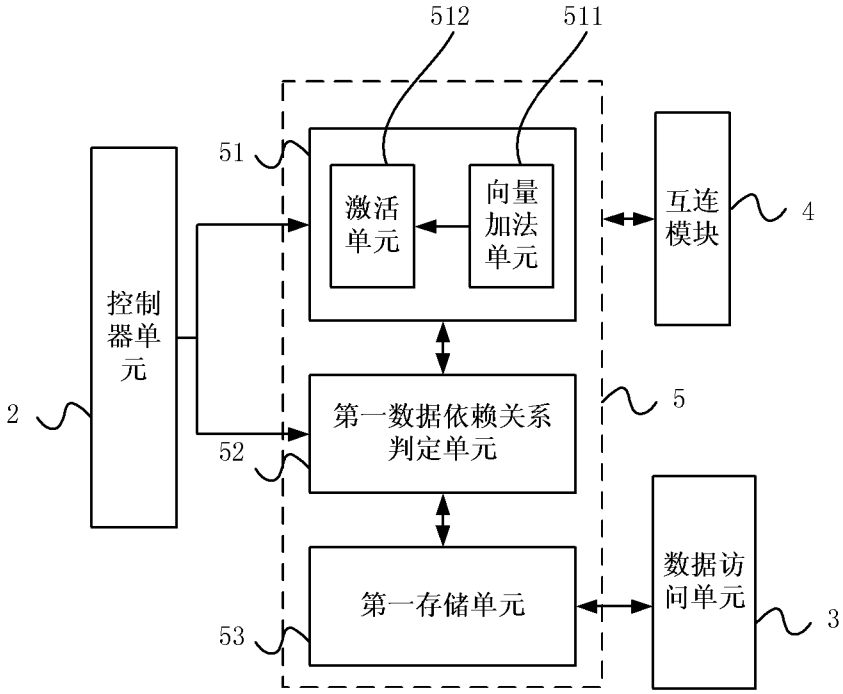
512
511
51
激活单元
向量加法单元
互连模块
4
控制器单元
2
第一数据依赖关系判定单元
5
52
数据访问单元
第一存储单元
53
3

图 5

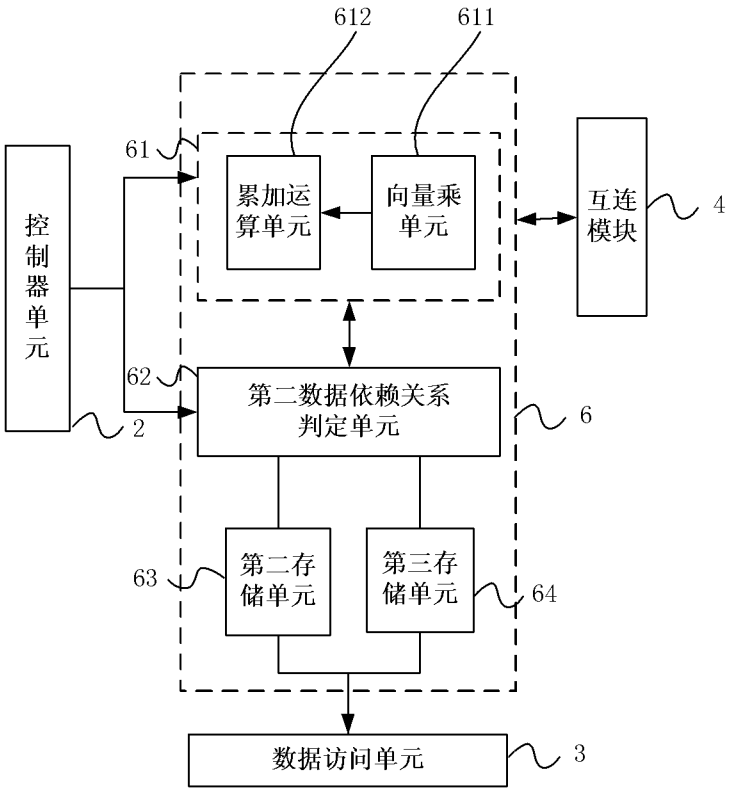
612
611
61
累加运算单元
向量乘单元
互连模块
4
控制器单元
62
第二数据依赖关系判定单元
2
6
63
第二存储单元
第三存储单元
64
数据访问单元
3

图 6

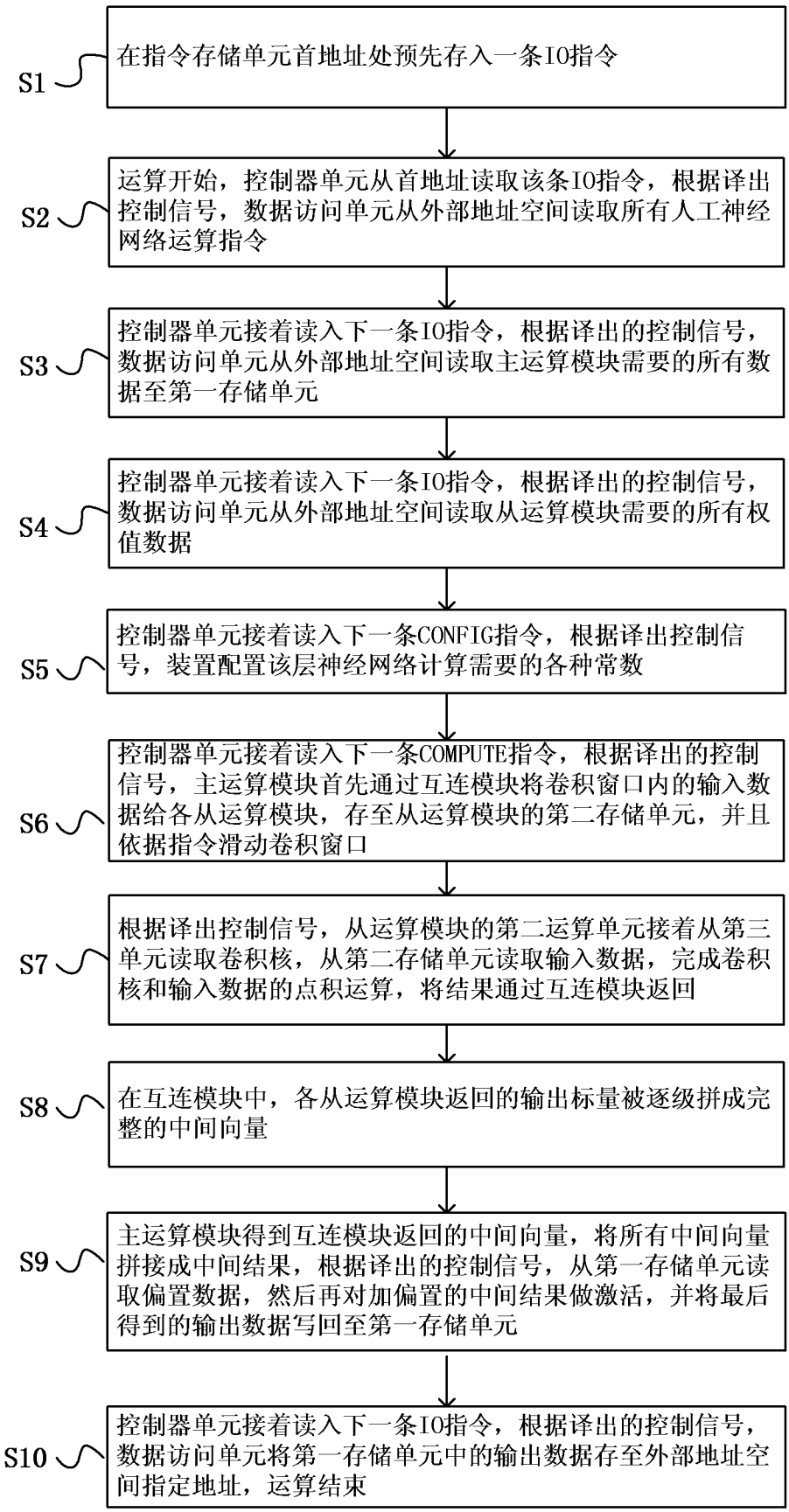
S1
在指令存储单元首地址处预先存入一条IO指令
S2
运算开始，控制器单元从首地址读取该条IO指令，根据译出控制信号，数据访问单元从外部地址空间读取所有人工神经网络运算指令
S3
控制器单元接着读入下一条IO指令，根据译出的控制信号，数据访问单元从外部地址空间读取主运算模块需要的所有数据至第一存储单元
S4
控制器单元接着读入下一条IO指令，根据译出的控制信号，数据访问单元从外部地址空间读取从运算模块需要的所有权值数据
S5
控制器单元接着读入下一条CONFIG指令，根据译出控制信号，装置配置该层神经网络计算需要的各种常数
S6
控制器单元接着读入下一条COMPUTE指令，根据译出的控制信号，主运算模块首先通过互连模块将卷积窗口内的输入数据给各从运算模块，存至从运算模块的第二存储单元，并且依据指令滑动卷积窗口
S7
根据译出控制信号，从运算模块的第二运算单元接着从第三单元读取卷积核，从第二存储单元读取输入数据，完成卷积核和输入数据的点积运算，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输出标量被逐级拼成完整的中间向量
S9
主运算模块得到互连模块返回的中间向量，将所有中间向量拼接成中间结果，根据译出的控制信号，从第一存储单元读取偏置数据，然后再对加偏置的中间结果做激活，并将最后得到的输出数据写回至第一存储单元
S10
控制器单元接着读入下一条IO指令，根据译出的控制信号，数据访问单元将第一存储单元中的输出数据存至外部地址空间指定地址，运算结束

图 7

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2016/080967

A. CLASSIFICATION OF SUBJECT MATTER

G06F 9/30 (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)

G06F, G06N

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)

WPL, EPODOC, CNPAT, CNKI, IEEE: neural network, forward, forward direction, master operation, slave operation, cooperate operation, convolutional, neural, network, CNN, filter, convolutional kernel, sigmoid, forward transmission, buffer, host, master, instruction

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 105184366 A (INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES), 23 December 2015 (23.12.2015), description, paragraphs 30-80	1-11
A	CN 104809426 A (NEC CORP.), 29 July 2015 (29.07.2015), the whole document	1-11
A	CN 103150596 A (BAIDU ONLINE NETWORK TECHNOLOGY (BEIJING) CO., LTD.), 12 June 2013 (12.06.2013), the whole document	1-11
A	US 5204938 A (LORAL AEROSPACE CORP.), 20 April 1993 (20.04.1993), the whole document	1-11

☐ 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
04 January 2017 (04.01.2017)	**26 January 2017 (26.01.2017)**
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 **LIU, Xu** Telephone No.: (86-10) **82245207**

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2016/080967

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 105184366 A	23 December 2015	None	
CN 104809426 A	29 July 2015	None	
CN 103150596 A	12 June 2013	None	
US 5204938 A	20 April 1993	None	

国际检索报告

国际申请号
PCT/CN2016/080967

A. 主题的分类

G06F 9/30(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)

G06F，G06N

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））

WPI，EPODOC，CNPAT，CNKI，IEEE：卷积，神经网络，卷积核，过滤器，前向，正向，缓存，主运算，从运算，协运算，指令，convolutional，neural，network，CNN，filter，convolutional kernel，sigmoid，forward transmission，buffer，host，master，instruction

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 105184366 A (中国科学院计算技术研究所) 2015年 12月 23日 (2015-12-23) 说明书第30-80段	1-11
A	CN 104809426 A (日本电气株式会社) 2015年 7月 29日 (2015-07-29) 全文	1-11
A	CN 103150596 A (百度在线网络技术北京有限公司) 2013年 6月 12日 (2013-06-12) 全文	1-11
A	US 5204938 A (LORAL AEROSPACE CORP.) 1993年 4月 20日 (1993-04-20) 全文	1-11

☐ 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 4日	2017年 1月 26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 刘栩 电话号码 (86-10)82245207

表 PCT/ISA/210 (第2页) (2009年7月)

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2016/080967

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 105184366 A	2015年 12月 23日	无	
CN 104809426 A	2015年 7月 29日	无	
CN 103150596 A	2013年 6月 12日	无	
US 5204938 A	1993年 4月 20日	无	