

(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
2018 年 2 月 8 日 (08.02.2018)



(10) 国际公布号
WO 2018/024232 A1

(51) 国际专利分类号:
G06N 3/063 (2006.01)

(21) 国际申请号: PCT/CN2017/095810

(22) 国际申请日: 2017 年 8 月 3 日 (03.08.2017)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201610635286.X 2016 年 8 月 5 日 (05.08.2016) CN

(71) 申请人: 上海寒武纪信息科技有限公司 (SHANGHAI CAMBRICON INFORMATION TECHNOLOGIES LTD.) [CN/CN]; 中国上海市浦东新区祖冲之路 2290 号 1 号楼 1004, Shanghai 201203 (CN)。

(72) 发明人: 陈云霁 (CHEN, Yunji); 中国上海市浦东新区祖冲之路 2290 号 1 号楼 1004, Shanghai 201203 (CN)。 刘少礼 (LIU, Shaoli); 中国上海市浦东新区祖冲之路 2290 号 1 号楼 1004, Shanghai 201203 (CN)。 韩栋 (HAN, Dong); 中国上海市浦东新区祖冲之路 2290 号 1 号楼 1004, Shanghai 201203 (CN)。 陈天石 (CHEN, Tianshi); 中国上海市浦东新区祖冲之路 2290 号 1 号楼 1004, Shanghai 201203 (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,

(54) Title: DEVICE AND METHOD FOR EXECUTING NEURAL NETWORK OPERATION

(54) 发明名称: 用于执行神经网络运算的装置及方法

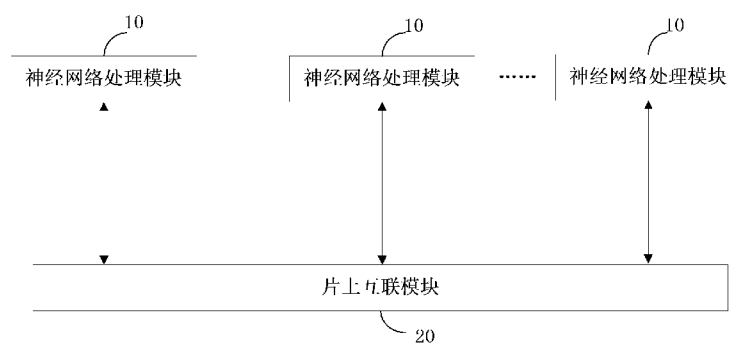


图 1

10 Neural network processing module
20 On-chip interconnection module

(57) Abstract: A device and method for executing a neural network operation. The device comprises: an on-chip interconnection module and a plurality of neural network processing modules connected thereto and communicating therewith. The neural network processing modules can read or write, via the on-chip interconnection module, data to or from another neural network processing module. For multi-core multi-layer artificial neural network operations, the neural network operations in each layer are divided for the plurality of neural network processing modules to execute the same, so as to obtain data of respective operation results. Further, the plurality of neural network processing modules exchange therebetween the data of respective operation results.

(57) 摘要: 一种用于执行神经网络运算的装置及方法, 该装置包括片上互联模块和与该片上互联单元通信连接的多个神经网络处理模块, 神经网络处理模块能够通过片上互联模块从其它神经网络处理模块中读写数据。在多核多层人工神经网络运算中, 要将每一层神经网络运算进行划分, 进而由多个神经网络处理模块进行运算, 得到各自的运算结果数据, 多个神经网络处理单元还将各自的运算结果数据进行数据交换。

CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, 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))。

用于执行神经网络运算的装置及方法

技术领域

本公开属于神经网络运算领域，尤其涉及一种用于执行神经网络运算的装置及方法。

5 背景技术

一种支持多核多层神经网络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。该方法的缺点之一是单个通用处理器的运算性能较低，无法满足通常的多核多层人工神经网络运算的性能需求。而多个通用处理器并行
10 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

另一种支持多核多层人工神经网络运算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
15 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，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。

20

发明内容

有鉴于此，本公开提供一种用于执行神经网络运算的装置及方法，能以低开销的方式实现一层或多层多核多层人工神经网络的运算，并且运算性能高效。

25 本公开提供一种用于执行神经网络运算的装置及方法，装置包括片上互联模块和与该片上互联单元通信连接的多个神经网络处理模块，神

神经网络处理模块能够通过片上互联模块从其它神经网络处理模块中读写数据。在多核多层人工神经网络运算中，要将每一层神经网络运算进行划分，进而由多个神经网络处理模块进行运算，得到各自的运算结果数据，多个神经网络处理单元还将各自的运算结果数据进行数据交换。

- 5 例如每进行一层神经网络计算之后，每个神经网络处理模块只计算得到了部分输出数据，而在执行下一层神经网络运算时，每个神经网络处理模块还会需求来自于其他神经网络处理模块的数据，因此每个神经网络处理的需要将自己计算得到的运算结果数据发送给对应的神经网络处理模块，用以使之计算下一层的神经网络。

- 10 用于执行神经网络运算的装置中，神经网络处理模块能够通过片上互联模块从其它神经网络处理模块中读写数据，还可从本地读写数据。

进一步，神经网络处理模块包括神经网络处理单元和高速存储单元；神经网络处理单元用于读写数据，高速存储单元用于存储本地数据。

- 15 进一步，用于执行神经网络运算的装置还包括外部存储模块，神经网络处理模块还能够通过片上互联单元从外部存储模块中读写数据。

进一步，神经网络处理单元包括指令队列、高速缓存单元、IO 读取单元和神经网络运算单元，其中：

- 20 进一步，指令队列存储有运算指令，IO 读取单元根据运算指令从该神经网络处理单元的外部读取数据，并将读取的数据缓存至高速缓存单元中，神经网络运算单元根据该运算指令从高速缓存单元中读取所缓存的数据，并执行神经网络运算，得到运算结果数据；

- 25 进一步，神经网络处理单元还包括同步关系单元，指令队列还存储有数据送达指令，神经网络运算单元向其它神经网络运算单元发送数据后，所述神经网络运算单元所对应的同步关系单元执行数据送达指令，以向其它神经网络运算单元所对应的同步关系单元发送一数据送达信号。

进一步，指令队列还存储有数据依赖指令，所述神经网络运算单元收到其其它神经网络运算单元所发送的数据后，其对应的同步关系单元执行所述数据依赖指令以检测是否是收到数据送达信号，若是，则继续

执行执行指令队列中的指令，否则阻塞指令队列。

进一步，指令队列还存储有数据同步指令，神经网络处理单元中的同步关系单元通过执行所述数据同步指令，以向其它神经网络处理单元中的同步关系单元发送一同步信号，以强制多个神经网络处理单元做同步操作。

进一步，片上互联模块包括一级互联模块和与该一级互联模块通信连接的多个二级互联模块，一级互联模块还与所述外部存储模块通信连接，所述多个二级互联模块与多个神经网络处理模块一一对应，其中，每个二级互联模块分别与相应神经网络处理模块中的神经网络处理单元和高速存储单元通信连接。

本公开还提供一种用于执行单层神经网络运算的方法，包括：

S1，多个神经网络处理模块中的每个神经网络处理模块直接从本地读取数据，和/或通过片上互联模块从其它神经网络处理模块中读取数据，其中，多个神经网络处理模块与该片上互联单元通信连接；

S2，每个神经网络处理模块根据读取的数据进行单层神经网络的部分运算，得到各自的运算结果数据；

S3，每个神经网络处理模块将各自的运算结果数据进行本地存储和/或通过片上互联模块将各自的运算结果数据写入至其他神经网络处理模块中。

进一步，步骤 S3 中，每个神经网络处理模块将各自的运算结果写入至其它神经网络处理模块中后，向其它特定的神经网络处理模块发送一数据送达信号。

本公开还提供一种用于执行多层神经网络运算的方法，对于每一层神经网络运算，执行上述步骤 S1-S3，并将该层各神经网络处理模块得到的运算结果数据用于下一层神经网络运算。

本公开所提供的用于执行神经网络运算的装置及方法，具有以下优点：

1、由于采用多核神经网络处理模块，允许单层的神经网络将任务划分，在多个神经网络处理模块上执行，并且采用了专用指令，允许在

执行多层神经网络的时候，可以在多个神经网络处理器之间相互传输计算得到的数据，故能实现多层多核的神经网络运算。

2、由于采用多核神经网络处理模块，解决了在执行多核多层神经网络处理运算的时候，单个处理器处理性能不足的问题，具有显著加速多核多层神经网络运算的效果。

3、由于采用了专用的数据指令，有效的解决了在执行多核多层神经网络时，其多个处理器之间需要交互大量数据的问题，具有显著加速多核多层神经网络运算的效果。

附图说明

- 10 图 1 是本公开提供的用于执行神经网络运算的装置的结构示意图；
图 2 是本公开中神经网络处理模块的结构示意图；
图 3 是本公开中外部存储模块的结构示意图；
图 4 是本公开中神经网络处理单元的结构示意图；
图 5 是本公开中片上互联单元的结构示意图；
15 图 6 是本公开执行一层全连接层运算实施例的流程图。

具体实施方式

本公开提供的用于执行人工神经网络运算的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。

- 25 图 1 是本公开提供的用于执行神经网络运算的装置的结构示意图，如图 1 所示，装置包括多个神经网络处理模块 10 及一个片上互联模块 20，多个神经网络处理模块 10 与该片上互联单元 20 通信连接，其中：

神经网络处理模块 10 能够通过片上互联模块 30 从其它神经网络处理模块 10 中读写数据，还可从本地读写数据。当要执行神经网络运算时，每个神经网络处理模块 10 作为一个核执行相应的运算，其运算所需的数据可直接从本地直接获取，也可通过片上互联模块 20 与其他神经网络处理模块 10 通信，以从其他神经网络处理模块 10 处读取运算所需的数据。各个神经网络处理模块 10 读取运算所需的数据后，执行相应的运算，得到各自的运算结果数据，在单层神经网络运算中，各个神经网络处理模块 10 可将各自的运算结果数据汇总至一个神经网络处理模块 10 中进行累加，以得到最终结果数据。在多层神经网络运算中，
5 当层各个神经网络处理模块 10 计算得到运算结果数据，可能在下一层作为运算所需的数据被其他神经网络处理模块 10 使用，这样在当层神经网络运算完毕后，各个神经网络处理模块 10 会进行数据交互，以准备进行下一层神经网络运算。

图 2 是本公开中神经网络处理模块的结构示意图，如图 2 所示，神经网络处理模块 10 包括神经网络处理单元 11 和高速存储单元 12；神经网络处理模块 10 在进行神经网络运算时，神经网络处理单元 11 直接与其对应的高速储存单元 12 中读取数据，和/或通过片上互联单元 20 从其它神经网络处理模块 10 中的神经网络处理单元 11 中读取数据，和/或通过片上互联单元 20 从其它神经网络处理模块 10 中的高速存储单元
15 12 中读取数据；每个神经网络处理模块 10 中的神经网络处理单元 11 根据读取的数据进行神经网络运算，得到各自的运算结果数据；在完成运算后，神经网络处理单元 11 将运算结果数据直接写入至与其对应的高速储存单元 12 中，和/或通过片上互联单元 20 将运算结果数据写入至其它神经网络处理模块 10 中的神经网络处理单元 11 中，和/或通过片上互
20 联单元 20 将运算结果数据写入至其它神经网络处理模块 10 中的高速存储单元 12 中。总之，神经网络处理单元 11 可直接从其对应的高速储存单元获取数据，也可以通过片上互联模块 20 获取其它位置的数据，这样避免了反复向内存读取数据，降低了内存访问带宽。

如图 3 所示，本公开提供的用于执行神经网络运算的装置还包括外

部存储模块 30，其与片上互联单元 20 通信连接，神经网络处理模块 10 还能够通过片上互联单元从外部存储模块中读写数据，利用外部存储模块 30，可以从外界向装置中导入新的数据，装置执行的最终执行结果数据也可以写入至外部存储模块 30，以供外部导出。其中，外部存储模块 30 可以通过硬件来实现（包括但不限于 FPGA、CGRA、专用集成电路 ASIC、模拟电路或忆阻器等）。

图 4 是本公开中神经网络处理单元 11 的结构示意图，如图 4 所示，神经网络处理单元 11 包括指令队列 111、神经网络运算单元 112、IO 读取单元 113、高速缓存单元 114 和同步关系单元 115。指令队列 111 存储有多种类型的指令，神经网络处理单元 11 根据不同的指令执行不同的操作。下表为各类指令的描述：

指令名称	操作码 1	操作码 2	操作码 3	操作码 4	操作码 5	……
ACK	0/1	0/1	0/1	0/1	0/1	……
FENCE	0/1	0/1	0/1	0/1	0/1	……
SYNC	0/1	0/1	0/1	0/1	0/1	……
COMPUTE	MLP	addr1	size1	addr2	size2	……
IO	src	dest	size			

指令包括指令名称以及多个操作码：

数据送达指令，指令名称为 ACK，其中各个操作码分别表示是否向该神经网络处理单元 11 发送数据送达信号（ACK 信号），神经网络处理单元 11 向其他神经网络处理单元 11 写入数据后，执行数据送达指令以发送数据送达信号给对应的神经网络处理单元 11，以表明数据已经传输到位；

数据依赖指令，指令名称为 FENCE，其中各操作码表示是否检查来自该神经网络处理单元 11 的 ACK 信号；神经网络处理单元 11 执行数据依赖指令以检测其所有依赖的数据是否已到达本神经网络处理单元。

数据同步指令，指令名称为 SYNC，其中各个操作码表示该神经网

络处理单元是否参与同步操作，神经网络处理单元 11 执行数据同步指令用以强制多个神经网络处理单元 11 做同步操作，即当多个神经网络都执行到当前指令后，这些神经网络处理单元才可以执行之后的指令；

运算指令，指令名称为 COMPUTE，其中第一个操作码表示具体的
5 计算任务，如 MLP，CONV，POOL 等，其余操作码用来表示输入输出数据的地址和大小，以及神经网络计算指令的配置信息。

输入输出指令，指令名称为 IO，其中的操作码分别表示搬运数据的起始地址，结束地址以及数据大小的信息，神经网络处理单元 11 执行输入输出指令以与其余模块之间进行通信数据。

10 IO 读取单元根据 113 根据指令队列 111 中的运算指令从该神经网络处理单元 11 的外部(如高速存储单元 12、其他神经网络处理单元 11 等)读取数据，并将读取的数据缓存至高速缓存单元 114 中，神经网络运算单元 112 根据该运算指令从高速缓存单元 114 中读取所缓存的数据，并执行神经网络运算，得到相应的运算结果数据；

15 神经网络运算单元 112 将运算结果数据写入至高速缓存单元 114 中，当需要将运算结果数据发送中外部(其他神经网络处理单元 11 等)时，IO 读取单元 113 从高速缓存单元 114 中读取运算结果数据，并将运算结果数据写入到该神经网络处理单元 11 的外部。

图 5 是本公开中片上互联单元的结构示意图，如图 5 所示，片上互
20 联 20 模块包括一级互联模块 21 和与该一级互联模块通信连接的多个二级互联模块 22，一级互联模块 21 还与外部存储模块 30 通信连接，多个二级互联模块 22 与多个神经网络处理模块 10 一一对应，其中，每个二级互联模块 22 分别与相应神经网络处理模块中的神经网络处理单元 11 和高速存储单元 12 通信连接。具体的，二级的互联模块 22 一个端口连
25 接神经网络处理单元 11，一个端口连接该神经网络处理单元对应的高速存储单元 12，另一个端口连接一级互联模块 21，一级互联模块 21 将多个二级互联模块 22 和外部存储模块 30 连接，用以保证这些模块之间的数据通路。这样，可以在保证各个神经网络处理单元 11 以及高速存储单元 12 和外部存储模块 30 之间相互通信，并且占用较小的面积开销。

另外，片上互联 20 模块还可以包括与每个二级互联模块 22 通信连接的一个或多个三级互联模块，还可以包括与每个三级互联模块通信连接的一个或多个四级互联模块，...，与每个 $n-1$ 级互联模块通信连接的一个或多个 n 级互联模块，其中 n 为大于等于 3 的正整数。每个 n 级互联模块分别与相应神经网络处理模块中的神经网络处理单元 11 和高速存储单元 12 通信连接。

采用本公开以上所描述的装置，可执行单层神经网络运算，包括：

S1，每个神经网络处理模块 10 根据其自身指令队列 11 中存储的计算指令，根据指令中操作码所指示的地址，直接从本地读取数据，和/或通过片上互联模块 20 从其它神经网络处理模块 10 中读取数据；

S2，每个神经网络处理模块 10 根据读取的数据进行单层神经网络的部分运算，得到各自的运算结果数据；

S3，每个神经网络处理模块 10 将各自的运算结果数据进行本地存储和/或通过片上互联模块 20 将各自的运算结果数据写入至其他神经网络处理模块 10 中。

对于多层神经网络运算，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，在下一层运算时，每个神经网络处理模块 10 根据新的运算指令从新的地址读取新的数据进行计算，并且依据新的指令在多核（即多个神经网络处理模块 10）之间分配计算任务。对于每一层神经网络运算，执行上述步骤 S1-S3，并将该层各神经网络处理模块 10 得到的运算结果数据用于下一层神经网络运算。

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，对本公开进一步详细说明。应当说明的是，虽然下面所述实施例仅讨论了全连接层的运算流程图，但是不局限在全连接层，也可以是所有的人工神经网络算法的运算流程图。

图 6 是本公开执行一层全连接层运算实施例的流程图，其执行过程如图 6 所示：

步骤 1：依据运算指令 COMPUTE，每个神经网络处理单元 11 从对应的高速存储单元 12 中读取数据，分别计算得到全连接层的部分运算

结果数据。

在每个神经网络处理单元 11 中，指令队列 111 将运算指令 COMPUTE 发送至神经网络运算单元 112 和 IO 读取单元 113，神经网络运算单元 112 根据运算指令 COMPUTE 中的指令名称，确定将要执行一层全连接层运算，具体的，IO 读取单元 113 根据运算指令 COMPUTE 中的地址从其对应的高速存储单元 12 中读取运算所需数据，并将读取的数据存储于高速缓存单元 114 中，神经网络运算单元 112 从高速缓存单元 114 中读取相应的数据，然后根据读取的数据执行运算指令 COMPUTE，以进行全连接层的部分运算，得到全连接层的部分运算结果数据作为输出数据。

步骤 2：依据输入输出指令 IO，每个神经网络处理单元 11 将自己计算的到的部分运算结果数据通过片上互联模块 20 发送给相应的神经网络处理单元 11。由于每个神经网络处理单元 11 只计算出部分运算结果数据，因此其需要将该部分输出数据发送给相应的神经网络处理单元 11 进行加和运算。

具体的，步骤 1 中神经网络运算单元 112 将计算得到的部分运算结果数据存储于高速缓存单元 114 中，指令队列 111 将输入输出指令 IO 发送给 IO 读取单元 113 后，IO 读取单元 113 执行输出指令 IO，以将存储于高速缓存单元 114 中的部分运算结果数据读取，并发送至外部的相应的神经网络处理单元 11。这里需要说明的是，每个神经网络处理单元 11 可能会将部分运算结果数据发送至一个对应的神经网络处理单元 11 中，也可能发送至多个对应的神经网络处理单元 11 中，也就是说，每个神经网络处理单元 11 也可能收到一个神经网络处理单元 11 发送的部分运算结果数据，也可能收到多个神经网络处理单元 11 发送的部分运算结果数据。

步骤 3：每个神经网络处理单元 11 将自己计算的到的部分运算结果数据发送给相应的神经网络处理单元 11 后，需要执行数据送达指令 ACK，以向对应的神经网络处理单元 11 发送数据送达信号。每个神经网络处理单元 11 需要向接受其发送数据的神经网络处理单元 11 发送数据送达

信号，用以表明其数据依赖关系。

步骤 4：依据数据依赖指令 FENCE，每个神经网络处理单元 11 检测其发送数据送达信号是否到达相应的神经网络处理单元 11，如果没有到达，则等待对应的数据送达信号到达相应的神经网络处理单元 11。对于每个将要进行加和运算神经网络处理单元 11，只有其收到所有其他神经网络处理单元 11 所发送的数据送达信号时，才表明其所需要的输入数据全部到达，从而执行加和运算。

步骤 5：依据运算指令 COMPUTE，每个的神经网络处理单元 11 汇集其他神经网络处理单元 11 的部分运算结果数据后，联合上自身运算所得的部分运算结果数据进行加和运算，得到最终的运算结果数据。

步骤 6：依据输入输出指令 IO，每个神经网络处理单元 11 将计算得到的最终的运算结果数据作为输出数据写入外部存储模块 30 中。在每个神经网络处理单元 11 中，将最终的运算结果数据写入外部存储模块 30 中的执行过程与步骤 2 类似，在此就不再赘述。

综上所述，本公开提供的装置和指令集，解决了 CPU 和 GPU 运算性能不足、前端译码开销大的问题，能有效支持多层人工神经网络运算，同时，针对多核多层人工神经网络运算采用专用片上存储，充分挖掘了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。

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。

权利要求

1、一种用于执行神经网络运算的装置，其特征在于，包括片上互联模块和与该片上互联单元通信连接的多个神经网络处理模块，其中：

5 所述神经网络处理模块能够通过所述片上互联模块从其它神经网络处理模块中读写数据。

2、根据权利要求 1 所述的用于执行多核多层神经网络运算的装置，其特征在于，所述神经网络处理模块还可从本地读写数据。

3、根据权利要求 2 所述的用于执行神经网络运算的装置，其特征在于，所述神经网络处理模块包括神经网络处理单元和高速存储单元；

10 所述神经网络处理单元用于读写数据，所述高速存储单元用于存储本地数据。

4、根据权利要求 1-3 任意一项所述的用于执行神经网络运算的装置，其特征在于，还包括外部存储模块，所述神经网络处理模块还能够通过所述片上互联单元从所述外部存储模块中读写数据。

15 5、根据权利要求 2 所述的用于执行神经网络运算的装置，其特征在于，所述神经网络处理单元包括指令队列、高速缓存单元、IO 读取单元和神经网络运算单元，其中：

所述指令队列存储有运算指令，所述 IO 读取单元根据所述运算指令从该神经网络处理单元的外部读取数据，并将读取的数据缓存至所述高速缓存单元中，所述神经网络运算单元根据该运算指令从所述高速缓存单元中读取所缓存的数据，并执行神经网络运算，得到运算结果数据；

20 所述神经网络运算单元将所述运算结果数据写入至所述高速缓存单元中，所述 IO 读取单元从所述高速缓存单元中读取所述运算结果数据，并将所述运算结果数据写入到该神经网络处理单元的外部。

25 6、根据权利要求 5 所述的用于执行神经网络运算的装置，其特征在于，所述神经网络处理单元还包括同步关系单元，所述指令队列还存储有数据送达指令，所述神经网络运算单元向其它神经网络运算单元发送数据后，所述神经网络运算单元所对应的同步关系单元执行数据送达

指令，以向其它神经网络运算单元所对应的同步关系单元发送一数据送达信号。

7、根据权利要求 6 所述的用于执行神经网络运算的装置，其特征在于，所述指令队列还存储有数据依赖指令，所述神经网络运算单元收到其其它神经网络运算单元所发送的数据后，其对应的同步关系单元执行所述数据依赖指令以检测是否是收到数据送达信号，若是，则继续执行执行指令队列中的指令，否则阻塞指令队列。

8、根据权利要求 6 所述的用于执行神经网络运算的装置，其特征在于，所述指令队列还存储有数据同步指令，神经网络处理单元中的同步关系单元通过执行所述数据同步指令，以向其它神经网络处理单元中的同步关系单元发送一同步信号，以强制多个神经网络处理单元做同步操作。

9、根据权利要求 3 所述的用于执行神经网络运算的装置，其特征在于，所述片上互联模块包括一级互联模块和与该一级互联模块通信连接的多个二级互联模块，所述一级互联模块还与所述外部存储模块通信连接，所述多个二级互联模块与所述多个神经网络处理模块一一对应，其中，每个二级互联模块分别与相应神经网络处理模块中的神经网络处理单元和高速存储单元通信连接。

10、一种用于执行单层神经网络运算的方法，其特征在于，包括：
S1，多个神经网络处理模块中的每个神经网络处理模块直接从本地读取数据，和/或通过片上互联模块从其它神经网络处理模块中读取数据，其中，多个神经网络处理模块与该片上互联单元通信连接；

S2，每个神经网络处理模块根据读取的数据进行单层神经网络的部分运算，得到各自的运算结果数据；

S3，每个神经网络处理模块将各自的运算结果数据进行本地存储和/或通过所述片上互联模块将各自的运算结果数据写入至其他神经网络处理模块中。

11、根据权利要求 10 所述的用于执行单层神经网络运算的方法，其特征在于，所述步骤 S3 中，每个神经网络处理模块将各自的运算结

果写入至其它神经网络处理模块中后，向其它特定的神经网络处理模块发送一数据送达信号。

- 12、一种用于执行多层神经网络运算的方法，其特征在于，对于每一层神经网络运算，执行如权利要求 10 所述的方法，并将该层各神经网络处理模块得到的运算结果数据用于下一层神经网络运算。
- 5

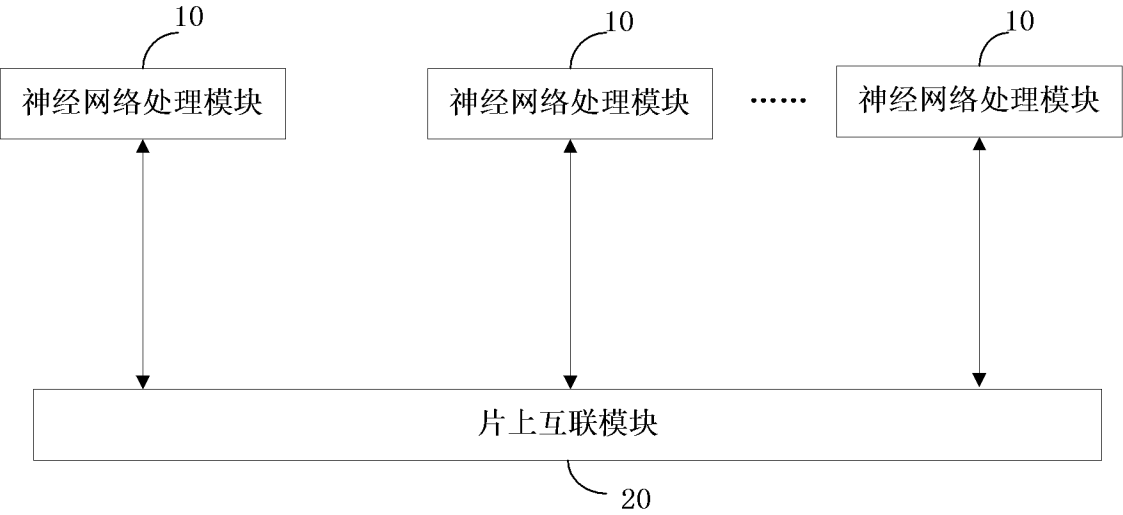


图 1

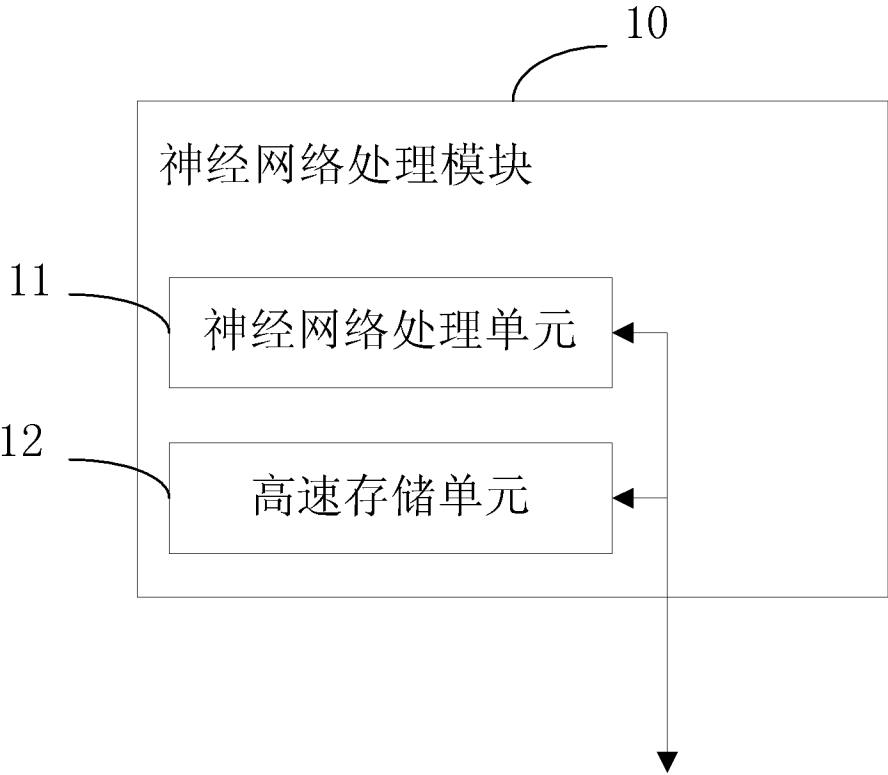


图 2

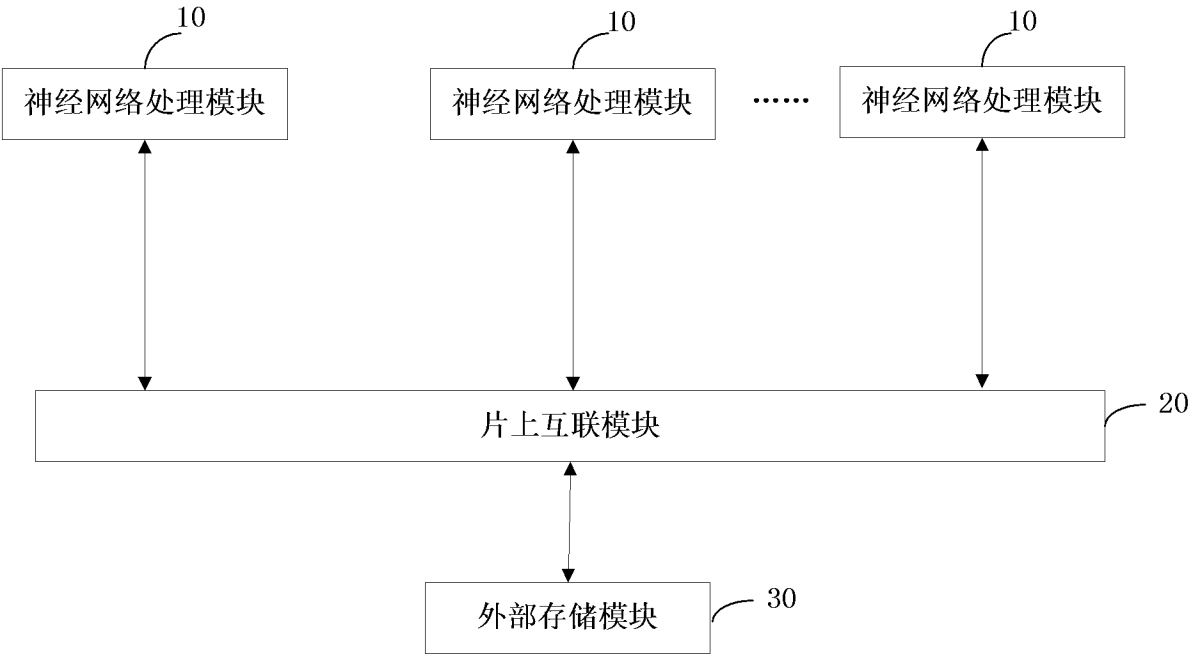


图 3

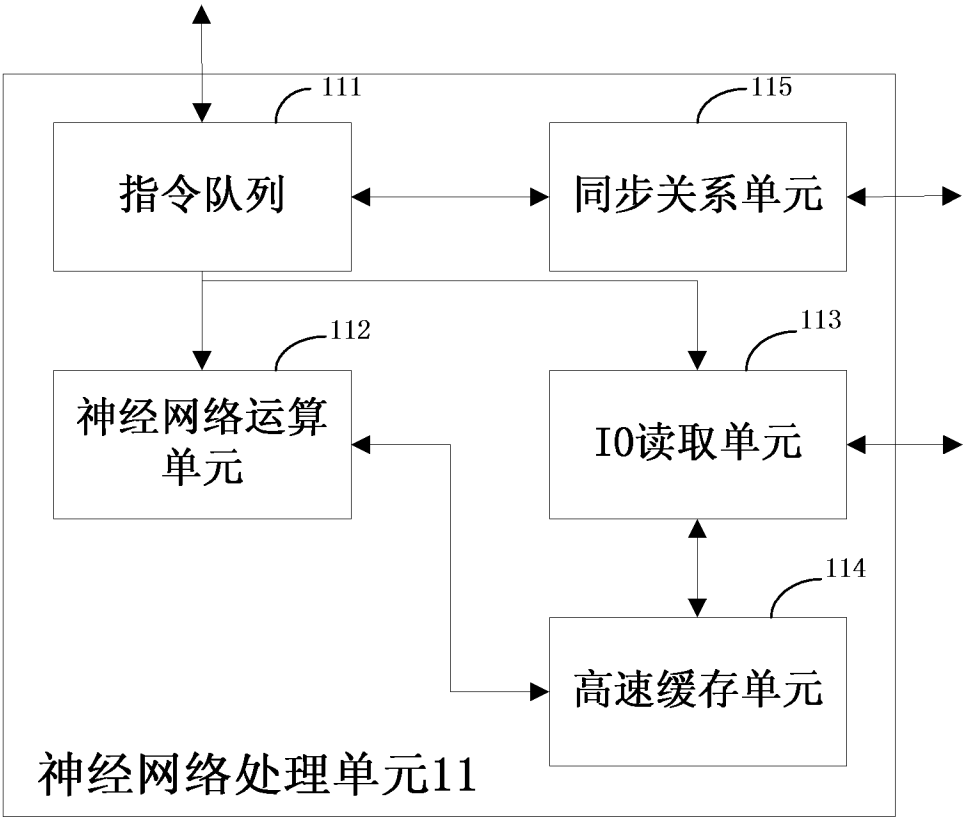


图 4

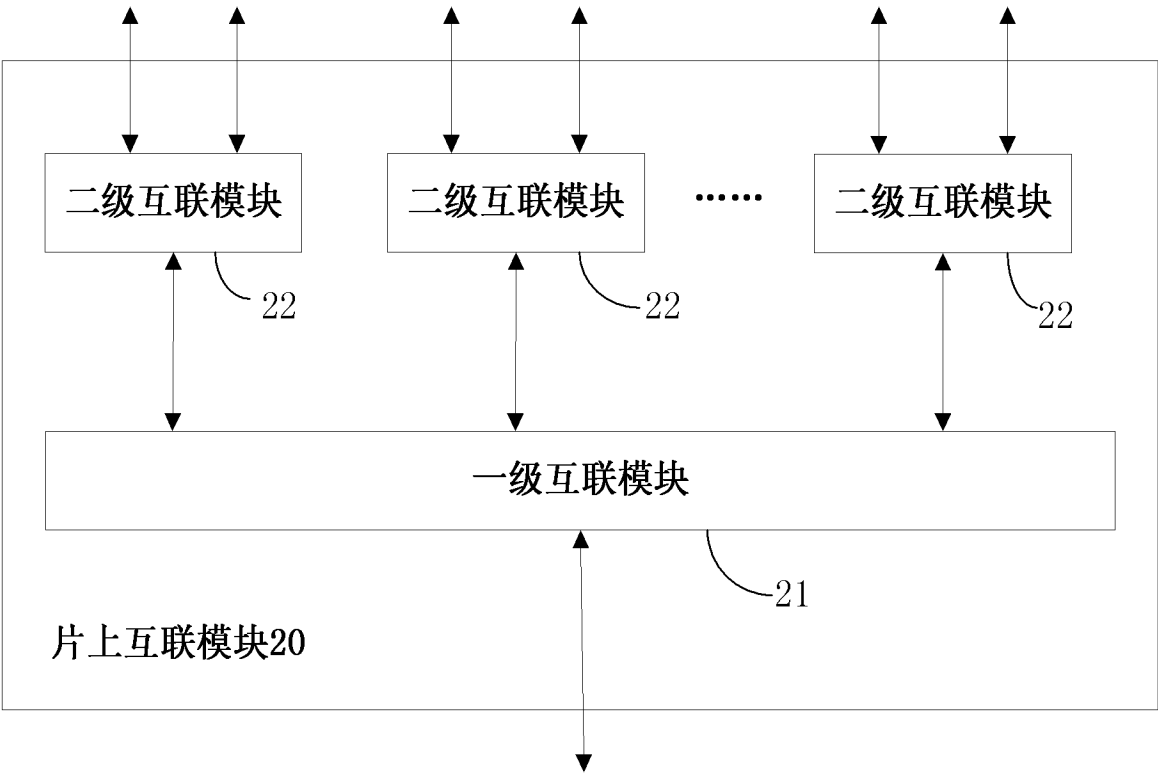


图 5

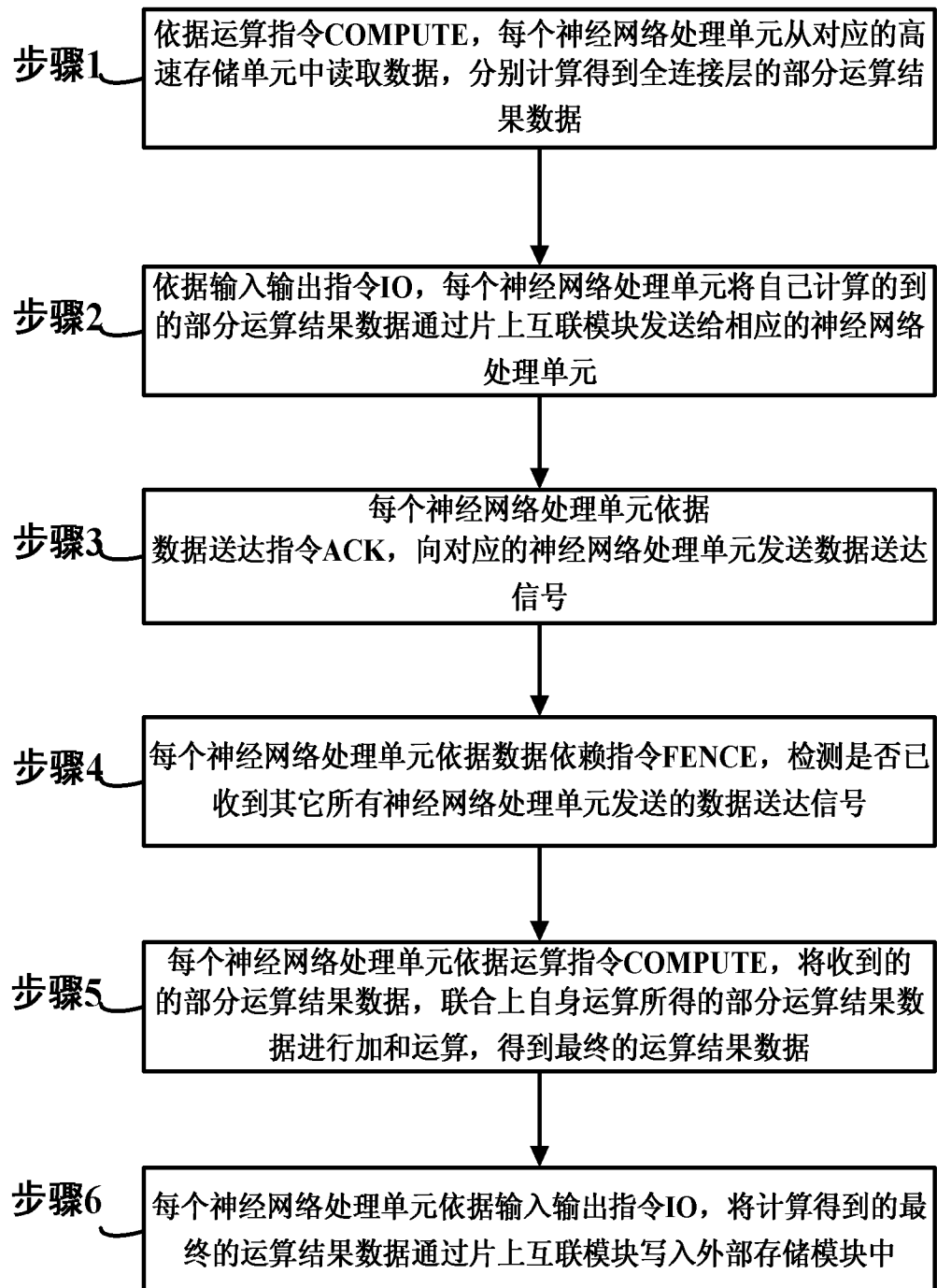


图 6

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2017/095810

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 G06F G05B

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)

WPI, EPODOC, CNPAT, CNKI, IEEE: neural network, operation, on-chip, neural, network, compute, SOC, system, on, chip, interconnection

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 105488565 A (INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES), 13 April 2016 (13.04.2016), claim 3, and description, paragraphs [0052]-[0081], and figure 1	1-12
A	CN 101882238 A (CHANG'AN UNIVERSITY), 10 November 2010 (10.11.2010), the whole document	1-12
A	CN 102193518 A (NANJING UNIVERSITY OF SCIENCE AND TECHNOLOGY), 21 September 2011 (21.09.2011), the whole document	1-12

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:	“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
“A” document defining the general state of the art which is not considered to be of particular relevance	“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
“E” earlier application or patent but published on or after the international filing date	“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
“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)	“&” document member of the same patent family
“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	

Date of the actual completion of the international search 22 October 2017 (22.10.2017)	Date of mailing of the international search report 07 November 2017 (07.11.2017)
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 JU, Bo Telephone No.: (86-10) 62413661

International application No.
PCT/CN2017/095810

Form PCT/ISA/210 (patent family annex) (July 2009)

国际检索报告

国际申请号

PCT/CN2017/095810

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 G06F G05B 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, EPDOC, CNPAT, CNKI, IEEE: 神经网络, 运算, 计算, 片上, 系统, 互联, neural, network, compute, SOC, system, on, chip, interconnection														
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>X</td> <td>CN 105488565 A (中国科学院计算技术研究所) 2016年 4月 13日 (2016 - 04 - 13) 权利要求3、说明书第[0052]-[0081]段、附图1</td> <td>1-12</td> </tr> <tr> <td>A</td> <td>CN 101882238 A (长安大学) 2010年 11月 10日 (2010 - 11 - 10) 全文</td> <td>1-12</td> </tr> <tr> <td>A</td> <td>CN 102193518 A (南京理工大学) 2011年 9月 21日 (2011 - 09 - 21) 全文</td> <td>1-12</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	X	CN 105488565 A (中国科学院计算技术研究所) 2016年 4月 13日 (2016 - 04 - 13) 权利要求3、说明书第[0052]-[0081]段、附图1	1-12	A	CN 101882238 A (长安大学) 2010年 11月 10日 (2010 - 11 - 10) 全文	1-12	A	CN 102193518 A (南京理工大学) 2011年 9月 21日 (2011 - 09 - 21) 全文	1-12
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求												
X	CN 105488565 A (中国科学院计算技术研究所) 2016年 4月 13日 (2016 - 04 - 13) 权利要求3、说明书第[0052]-[0081]段、附图1	1-12												
A	CN 101882238 A (长安大学) 2010年 11月 10日 (2010 - 11 - 10) 全文	1-12												
A	CN 102193518 A (南京理工大学) 2011年 9月 21日 (2011 - 09 - 21) 全文	1-12												
☐ 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 2017年 10月 22日		国际检索报告邮寄日期 2017年 11月 7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		授权官员 鞠博 电话号码 (86-10)62413661												

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2017/095810

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	105488565	A	2016年 4月 13日	WO 2017084330 A1	2017年 5月 26日
CN	101882238	A	2010年 11月 10日	无	
CN	102193518	A	2011年 9月 21日	无	