(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 10 月 25 日 (25.10.2018)

WIPO | PCT

(10) 国际公布号
WO 2018/192492 A1

(51) 国际专利分类号:
G06N 3/08 (2006.01)

(21) 国际申请号: PCT/CN2018/083379

(22) 国际申请日: 2018 年 4 月 17 日 (17.04.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201710261742.3 2017年4月20日 (20.04.2017) CN
201710279834.4 2017年4月25日 (25.04.2017) CN
201710279655.0 2017年4月25日 (25.04.2017) CN

(71) 申请人: 上海寒武纪信息科技有限公司 (SHANGHAI CAMBRICON INFORMATION TECHNOLOGY CO., LTD) [CN/CN]; 中国上海市浦东新区同汇路 168 号 B 座 6 层, Shanghai 201306 (CN)。

(72) 发明人: 陈天石 (CHEN, Tianshi); 中国上海市浦东新区同汇路 168 号 B 座 6 层, Shanghai 201306 (CN)。 庄毅敏 (ZHUANG, Yimin); 中国上海市浦东新区同汇路 168 号 B 座 6 层, Shanghai 201306 (CN)。 刘道福 (LIU, Daofu); 中国上海市浦东新区同汇路 168 号 B 座 6 层, Shanghai 201306 (CN)。 陈小兵 (CHEN, Xiaobing); 中国上海市浦东新区同汇路 168 号 B 座 6 层, Shanghai 201306 (CN)。 王在 (WANG, Zai); 中国上海市浦东新区同汇路 168 号 B 座 6 层, Shanghai 201306 (CN)。 刘少礼 (LIU, Shaoli); 中国上海市浦东新区同汇路 168 号 B 座 6 层, Shanghai 201306 (CN)。

(74) 代理人: 广州三环专利商标代理有限公司 (SCIHEAD IP LAW FIRM); 中国广东省广州市

(54) **Title:** COMPUTING APPARATUS AND RELATED PRODUCT

(54) 发明名称: 一种运算装置及相关产品

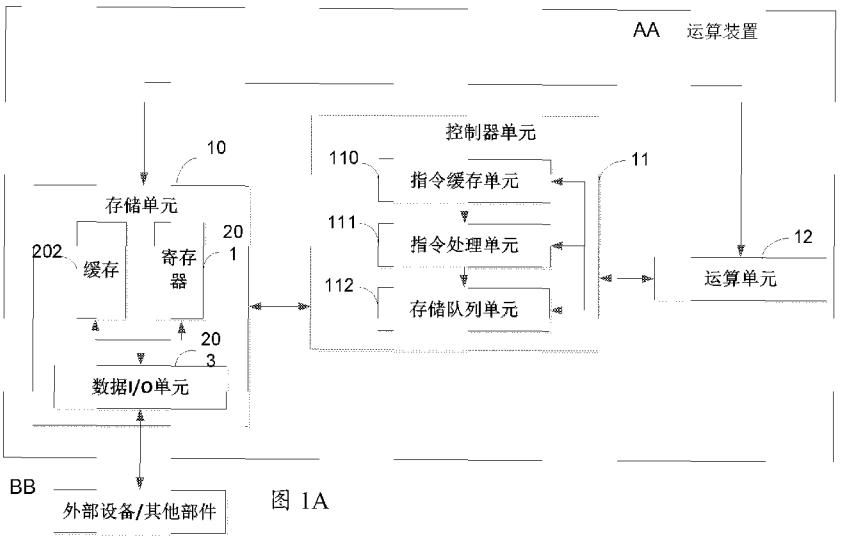


图 1A

10	STORAGE UNIT	112	STORAGE QUEUE UNIT
11	CONTROLLER UNIT	201	REGISTER
12	COMPUTING UNIT	202	CACHE
110	INSTRUCTION CACHE UNIT	203	DATA I/O UNIT
111	INSTRUCTION PROCESSING UNIT	AA	COMPUTING APPARATUS
		BB	EXTERNAL DEVICE/OTHER COMPONENTS

(57) **Abstract:** The present application provides a computing apparatus and a related product. The computing apparatus is used for performing calculation of a network model; the network model comprises a neural network model and/or a non-neural network model; the computing apparatus comprises a computing unit, a controller unit, and a storage unit; the storage unit comprises a data input/output unit, a storage medium, and a scalar data storage unit. The technical solution provided by the present application provides a fast calculation speed and can save energy.

WO 2018/192492 A1

[见续页]

越秀区先烈中路80号汇华商贸大厦1508室, Guangdong 510070 (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, 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))。

(57) 摘要：本申请提供了一种运算装置及相关产品，所述运算装置用于执行网络模型的计算，所述网络模型包括：神经网络模型和/或非神经网络模型；所述运算装置包括：运算单元、控制器单元以及存储单元，所述存储单元包括：数据输入输出单元、存储介质和标量数据存储单元。本申请提供的技术方案具有计算速度快，节能的优点。

一种运算装置及相关产品

技术领域

本申请涉及人工智能技术领域，具体涉及一种运算装置及相关产品。

背景技术

深度学习的概念源于人工神经网络的研究。含多隐层的多层感知器就是一种深度学习结构。深度学习通过组合低层特征形成更加抽象的高层表示属性类别或特征，以发现数据的分布式特征表示。

在实践中发现，现有的深度学习系统通常需要占用大量存储资源及运算资源，尤其对于复杂运算，大大降低了系统运算效率。因此，如何降低深度学习中存储资深及运算资源消耗的问题亟待解决。

申请内容

本申请实施例提供了一种运算装置及相关产品，可降低深度学习中存储资深及运算资源消耗。

本申请实施例第一方面提供了一种运算装置，所述运算装置包括存储单元、运算单元和控制器单元，其中，

所述存储单元，用于存储数据和指令;

所述控制器单元，用于从所述存储单元中提取第一指令以及所述第一指令对应的第一数据，所述第一数据包括输入神经元数据和权值数据，所述第一指令包括排序指令或者稀疏处理指令;

所述运算单元，用于响应所述第一指令，对所述输入神经元数据和所述权值数据执行所述第一指令对应的运算操作，得到运算结果。

第二方面，本申请实施例提供了一种运算方法，应用于运算装置，所述运算装置包括存储单元、运算单元和控制器单元，其中，

所述存储单元存储数据和指令;

所述控制器单元从所述存储单元中提取第一指令以及所述第一指令对应

的第一数据，所述第一数据包括输入神经元数据和权值数据，所述第一指令包括排序指令或者稀疏处理指令；

所述运算单元响应所述第一指令，对所述输入神经元数据和所述权值数据执行所述第一指令对应的运算操作，得到运算结果。

第三方面，本申请实施例提供了一种神经网络计算装置，该神经网络计算装置包括一个或者多个第一方面所述的运算装置。该神经网络计算装置用于从其他处理装置中获取待运算数据和控制信息，并执行指定的神经网络运算，将执行结果通过 I/O 接口传递给其他处理装置；

当所述神经网络计算装置包含多个所述运算装置时，所述多个所述运算装置间可以通过特定的结构进行连接并传输数据；

其中，多个所述运算装置通过快速外部设备互连总线（Peripheral Component Interconnect-Express，PCI-E 或 PCIe）PCIE 总线进行互联并传输数据，以支持更大规模的神经网络的运算；多个所述运算装置共享同一控制系统或拥有各自的控制系统；多个所述运算装置共享内存或者拥有各自的内存；多个所述运算装置的互联方式是任意互联拓扑。

第四方面，本申请实施例提供了一种组合处理装置，该组合处理装置包括如第一方面所述的运算装置、通用互联接口，和其他处理装置。该神经网络计算装置与上述其他处理装置进行交互，共同完成用户指定的操作。

第五方面，本申请实施例提供了一种神经网络芯片，该神经网络芯片包括上述第一方面所述的运算装置、上述第三方面所述的神经网络计算装置或者上述第四方面所述的组合处理装置。

第六方面，本申请实施例提供了一种神经网络芯片封装结构，该神经网络芯片封装结构包括上述第五方面所述的神经网络芯片；

第七方面，本申请实施例提供了一种板卡，该板卡包括上述第六方面所述的神经网络芯片封装结构。

第八方面，本申请实施例提供了一种电子装置，该电子装置包括上述第六方面所述的神经网络芯片或者上述第七方面所述的板卡。

可以看出，在本申请实施例的方案中，存储单元存储数据和指令，控制器单元从存储单元中提取第一指令以及第一指令对应的第一数据，第一数据包括

输入神经元数据和权值数据，第一指令为排序指令或者稀疏处理指令，运算单元响应第一指令，对输入神经元数据和权值数据执行第一指令对应的运算操作，得到运算结果，可降低深度学习中存储资深及运算资源消耗，提高了计算效率。

另外，在一些实施例中，所述电子装置包括数据处理装置、机器人、电脑、打印机、扫描仪、平板电脑、智能终端、手机、行车记录仪、导航仪、传感器、摄像头、服务器、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、交通工具、家用电器、和/或医疗设备。

在一些实施例中，所述交通工具包括飞机、轮船和/或车辆；所述家用电器包括电视、空调、微波炉、冰箱、电饭煲、加湿器、洗衣机、电灯、燃气灶、油烟机；所述医疗设备包括核磁共振仪、B 超仪和/或心电图仪。

附图说明

为了更清楚地说明本申请实施例中的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

图 1A 为本申请实施例提供的一种运算装置的结构示意图；

图 1B 为本申请实施例提供的一种运算装置的另一结构示意图；

图 1C 为本申请实施例提供的稀疏模式 1 的处理过程示意图；

图 1D 为本申请实施例提供的稀疏模式 2 的处理过程示意图；

图 1E 为本申请实施例提供的稀疏模式 3 的处理过程示意图；

图 1F 为本申请实施例提供的运算单元及其连接关系的结构示意图；

图 1G 为本申请实施例提供的第 1 个向量归并单元的结构示意图；

图 1H 为本申请实施例提供的一种运算装置的另一结构示意图；

图 1I 为本申请实施例提供的一种运算装置的另一结构示意图；

图 1J 为本申请实施例提供的主处理电路的结构示意图；

图 1K 为本申请实施例提供的神经网络模型的结构图的示意图；

图 1L 为本申请实施例提供的一种运算装置的另一结构示意图；

图 2A 为本申请实施例提供的一种组合处理装置的结构示意图；

图 2B 为本申请实施例提供的另一种组合处理装置的结构示意图。

具体实施方式

下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

本申请的说明书和权利要求书及所述附图中的术语“第一”、“第二”、“第三”和“第四”等是用于区别不同对象，而不是用于描述特定顺序。此外，术语“包括”和“具有”以及它们任何变形，意图在于覆盖不排他的包含。例如包含了一系列步骤或单元的过程、方法、系统、产品或设备没有限定于已列出的步骤或单元，而是可选地还包括没有列出的步骤或单元，或可选地还包括对于这些过程、方法、产品或设备固有的其它步骤或单元。

在本文中提及“实施例”意味着，结合实施例描述的特定特征、结构或特性可以包含在本申请的至少一个实施例中。在说明书中的各个位置出现该短语并不一定均是指相同的实施例，也不是与其它实施例互斥的独立的或备选的实施例。本领域技术人员显式地和隐式地理解的是，本文所描述的实施例可以与其它实施例相结合。

首先介绍本申请使用的运算装置。参阅图 1A，提供了一种运算装置，该运算装置包括：存储单元 10、控制器单元 11 和运算单元 12，其中，控制器单元 11 与存储单元 10 以及运算单元 12；其中，

该存储单元 10 可以包括：数据输入输出单元（数据 I/O 单元）203，

数据输入输出单元 203，用于获取输入数据、权值数据、网络模型以及计算指令；

控制器单元 11，用于从所述存储单元提取第一指令，解析该第一指令得到该计算指令的操作码以及操作域，提取该操作域对应的输入数据以及权值数据，将该操作码、输入数据以及权值数据发送给所述运算单元，所述操作码包括以下至少一种：矩阵计算指令的操作码、向量计算指令操作码、激活计算指

令操作码、偏置计算指令操作码、卷积计算指令操作码、转换计算指令操作码等等;

运算单元 12，用于依据该操作码对该输入数据以及权值数据执行该操作码对应的运算得到第一指令的结果。

可选地，该控制器单元包括：指令缓存单元 110、指令处理单元 111 和存储队列单元 113，所述指令缓存单元 110 用于对指令进行缓存，所述指令处理单元 111 用于实现译码功能;

指令缓存单元 110，用于缓存所述第一指令;

指令处理单元 111，用于解析所述第一指令得到该第一指令的操作码以及操作域;

存储队列单元 113，用于存储指令队列，所述指令队列包括：按该队列的前后顺序待执行的多个计算指令或操作码。

该计算指令可以包括：一个或多个操作域以及一个操作码。该计算指令可以包括神经网络运算指令。以神经网络运算指令为例，如表 1 所示，其中，寄存器号 0、寄存器号 1、寄存器号 2、寄存器号 3、寄存器号 4 可以为操作域。其中，每个寄存器号 0、寄存器号 1、寄存器号 2、寄存器号 3、寄存器号 4 可以是一个或者多个寄存器的号码。

操作码	寄存器号 0	寄存器号 1	寄存器号 2	寄存器号 3	寄存器号 4
COMPUTE	输入数据起始地址	输入数据长度	权值起始地址	权值长度	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		

可选的，存储单元还可以包括：寄存器 201 和缓存 202 和数据 I/O 单元。

存储介质 201 可以为片外存储器，当然在实际应用中，也可以为片内存储

器，用于存储数据块，该数据块具体可以为 n 维数据，n 为大于等于 1 的整数，例如，n=1 时，为 1 维数据，即向量，如 n=2 时，为 2 维数据，即矩阵，如 n=3 或 3 以上时，为多维张量。

可选地，上述第一指令可以为向量指令，向量指令可以为以下至少一种：向量加法指令（VA）、向量加标量指令（VAS）、向量减法指令（VS）、向量乘法指令（VMV）、向量乘标量指令（VMS）、向量除法指令（VD）、标量除向量指令（SDV）、向量间与指令（VAV）、向量内与指令（VAND）、向量间或指令（VOV）、向量内或指令（VOR）、向量指数指令（VE）、向量对数指令（VL）、向量大于判定指令（VGT）、向量等于判定指令（VEQ）、向量非指令（VINV）、向量选择合并指令（VMER）、向量最大值指令（VMAX）、标量扩展指令（STV）、标量替换向量指令（STVPN）、向量替换标量指令（VPNTS）、向量检索指令（VR）、向量点积指令（VP）、随机向量指令（RV）、循环移位指令（VCS）、向量加载指令（VLOAD）、向量存储指令（VS）、向量搬运指令（VMOVE）、矩阵乘向量指令（MMV）、向量乘矩阵指令（VMM）、矩阵乘标量指令（VMS）、张量运算指令（TENS）、矩阵加法指令（MA）、矩阵减法指令（MS）、矩阵检索指令（MR）、矩阵加载指令（ML）、矩阵存储指令（MS）、矩阵搬运指令（MMOVE）。

可选地，如图 1B 所示，图 1B 为图 1A 所描述的运算装置的一种变型结构，其与图 1A 相比较，还可以包括：配置解析单元 13、映射单元 14 和稀疏单元 15，具体如下：

所述存储单元 10，用于存储数据和指令；

所述控制器单元 11，用于从所述存储单元中提取第一指令以及所述第一指令对应的第一数据，所述第一数据包括输入神经元数据和权值数据，所述第一指令包括排序指令或者稀疏处理指令；

所述运算单元 12，用于响应所述第一指令，对所述输入神经元数据和所述权值数据执行所述第一指令对应的运算操作，得到运算结果。

其中，运算单元 12 可支持多种数据类型的运算，根据指令要求选择相应的运算器完成对应运算，例如，数据类型可以为 16 位定点数据或者 32 位浮点数据等。举例说明下，指令是矩阵加矩阵，选择加法器；指令是矩阵乘矩阵，

选择乘法器和加法器，指令是16位定点运算指令，接收该指令进行16位定点运算，等等。

其中，上述数据可以包括以下至少一种数据类型：整型数据、离散型数据、连续型数据、幂次型数据、浮点型数据或者定点型数据，数据表示的长度可为32位长度浮点数据，16位长度定点数据等等；数据可包括以下至少一种：输入神经元数据、权值数据和偏置数据。

可选地，在所述第一指令为所述稀疏处理指令以及所述第一数据还包括预设配置数据时，其中，

所述配置解析单元13，用于根据所述预设配置数据设置映射模式；

所述映射单元 14，用于根据所述映射模式对所述输入神经元和所述权值数据进行映射处理，得到输入神经元-权值对，所述输入神经元-权值对为映射处理后的输入神经元数据与权值数据之间的映射关系；

所述指令缓存单元110，用于接收由所述控制器单元发送的目标指令；

所述指令处理单元111，用于将所述目标指令译码为运算指令；由所述运算单元对所述输入神经元-权值对执行运算操作，得到运算结果。

其中，预设配置数据可包括以下至少一种：数据类型，或者，稀疏参数。目标指令为矩阵乘矩阵计算指令，其译码后得到运算指令，该运算指令可以包括乘法运算指令和加法运算指令。第一指令包括至少一个目标指令。

可选地，稀疏单元 15，用于依据所述稀疏参数对所述运算结果进行稀疏处理，得到稀疏处理后的运算结果。

可选地，所述稀疏参数包括稀疏模式；所述映射单元13根据所述映射模式对所述输入神经元和所述权值进行映射处理，具体为：

当处于所述稀疏模式为第一稀疏模式时，获取所述第一稀疏模式对应的权值稀疏序列，并依据该权值稀疏序列对所述权值进行映射处理；

当处于所述稀疏模式为第二稀疏模式时，获取所述第二稀疏模式对应的神经元稀疏序列，并依据该神经元稀疏序列对所述输入神经元进行映射处理；

当处于所述稀疏模式为第三稀疏模式时，获取所述第三稀疏模式对应的权值稀疏序列和神经元稀疏序列，并依据该权值稀疏序列和神经元稀疏序列对所述输入神经元和所述权值数据进行映射处理。

可选地，稀疏参数可包括以下至少一种：稀疏标志，稀疏率，稀疏模式等。其中，稀疏标志用于确定是否进行稀疏处理，例如，可以用 0 表示不进行稀疏处理，1 表示进行稀疏处理，也可以用 1 表示不进行稀疏处理，0 表示进行稀疏处理。可以理解，还可以用根据需求灵活选择稀疏标志的表示方式。在进行稀疏处理的情况下，稀疏率表示每次进行稀疏处理神经元数据和/或权值数据的比例，例如 5%，10%，25%等等。稀疏模式表示稀疏处理的具体模式，本申请实施例中，稀疏模式主要至少包括 3 种：稀疏模式 1，仅权值稀疏处理；稀疏模式 2，仅神经元数据稀疏处理；稀疏模式 3，权值和神经元数据均稀疏处理，当然，稀疏模式还可以为以上至少两种模式组合，例如，稀疏模式 1+稀疏模式 2。另外，在未作稀疏处理的情况下，对应的稀疏模式记作模式 0。又例如，神经元数据的稀疏率和权值数据的稀疏率还可不一样，例如，本申请实施例提供一种稀疏率的表示方式（A，B），其中，A 为神经元数据的稀疏率，B 为权值数据的稀疏率，例如为（5%，6%），即神经元数据的稀疏率为 5%，权值数据的稀疏率为 6%。

可选地，在没有稀疏标志的情况下，至少包括以下四种稀疏模式：稀疏模式 0，不做稀疏；稀疏模式 1，仅权值数据稀疏；稀疏模式 2，仅神经元数据稀疏；稀疏模式 3，权值和神经元数据都稀疏。

举例来说，上一层的输出神经元数据作为下一层的输入神经元数据时候，因为输出神经元数据已作稀疏处理，所以在下一层运算里，假如，稀疏的标准不改变，输入神经元数据就不需要重复做稀疏。

其中，所述配置解析单元 13 由神经网络的配置数据解析获得的稀疏模式进而设置映射单元的处理模式，即根据不同的稀疏模式对应不同的映射模式。可选地，存储单元中预先存储稀疏模式与映射模式之间的映射关系，进而，依据该映射关系确定与稀疏模式对应的映射模式，在不同的映射模式下，根据神经元稀疏序列和权值稀疏序列做稀疏化，当然，映射关系不一定存储在存储单元中，还可以存储在片外存储器上，或者，还可以存储在其他设备（具备存储器功能的电子装置）上。存储单元中可预先存储权值稀疏序列和/或神经元稀疏序列。

可选地，所述稀疏单元 15 依据所述稀疏参数对所述运算结果进行稀疏处理，具体为：

对神经元数据的元素绝对值排序，根据稀疏率计算获得需要稀疏的元素个数，根据需要稀疏的元素个数对排序后的神经元数据的元素作稀疏处理，并将稀疏后的稀疏神经元数据和神经元稀疏序列发送至所述控制器单元 11。

其中，可对输出神经元的元素绝对值进行排序，根据稀疏率计算获得需要稀疏的元素个数，然后对输出神经元的元素绝对值小于预设阈值的元素作稀疏处理，即置其值为 0，预设阈值可由用户自行设置或者系统默认，稀疏率可动态调整。

可选地，所述稀疏单元 15 依据所述稀疏参数对所述运算结果进行稀疏处理，具体为：

神经元数据为 0 的元素保持 0 不变，神经元数据在预设取值区间内的元素置为 0。

其中，神经元数据为 0 的元素保持 0 不变，神经元数据在预设取值区间内的元素置为 0 值，预设取值空间可以由用户自行设置或者系统默认。

举例说明下，如图 1C 所示，图 1C 为稀疏模式 1 的示例图。其中，稀疏权值数据只包括权值中非零的数据只存储 w1,w5,w8,w9 的权值数据，权值稀疏序列用于索引稀疏权值，如权值稀疏序列为 100010011 表示 w1,w5,w8,w9 的权值为非零值，而 w2,w3,w4,w6,w7 为 0。稀疏序列的表示方式并不唯一，可以使用 0 表示非稀疏，即数值非零，用 1 表示稀疏，即数值为零，也可以采用其他可行方式。根据权值稀疏序列，选择对应的输入神经元数据，如图 1C 中选择 d1,d5,d8,d9 输入神经元数据，通过对输入神经元数据和权值数据的映射，获得对应的输入神经元-权值对。

再举例说明下，如图 1D 所示，图 1D 为本申请实施例提供的稀疏模式 2 的示例图，其中稀疏神经元数据只包括神经元中非零的数据，如图 1D 只存储 d1,d3,d5,d8 的神经元数据，神经元稀疏序列用于索引稀疏神经元数据，如神经元稀疏序列 101010010 表示 d1,d3,d5,d8 的神经元为非零值，而 d2,d4,d6,d7,d9 为 0。应当认识到，稀疏序列的表示方式并不唯一，可以使用 0 表示非稀疏，即数值非零，用 1 表示稀疏，即数值为零，也可以采用其他可行方式。根据神

经元稀疏序列，选择对应的权值数据，如图 1D 中选择 w1,w3,w5,w8 的权值数据，通过对输入神经元数据和权值数据进行映射处理，得到对应的输入神经元-权值对。

再举例说明下，如图 1E 所示，图 1E 为本申请实施例提供的稀疏模式 3 的示例图，即根据神经元稀疏序列和权值稀疏序列，选择为非零值的输入神经元数据和权值数据，如图 1E 所示，选择神经元数据 d1,d5,d8 和权值数据 w1,w5,w8，通过对输入神经元数据和权值数据的映射处理，得到对应的输入神经元-权值对。

可选地，基于上述运算装置，可以实现如下运算方法，具体如下

所述存储单元 10 存储数据和指令;

所述控制器单元 11 从所述存储单元 10 中提取第一指令以及所述第一指令对应的第一数据，所述第一数据包括输入神经元数据和权值数据，所述第一指令包括排序指令或者稀疏处理指令;

所述运算单元 12 响应所述第一指令，对所述输入神经元数据和所述权值数据执行所述第一指令对应的运算操作，得到运算结果。

进一步可选地，所述控制器单元 11 包括：指令缓存单元 110 和指令处理单元 111。

进一步可选地，在所述第一指令为所述稀疏处理指令以及所述第一数据还包括预设配置数据时，其中，

所述配置解析单元 13 根据所述预设配置数据设置映射模式;

所述映射单元 14 根据所述映射模式对所述输入神经元和所述权值数据进行映射处理，得到输入神经元-权值对，所述输入神经元-权值对为映射处理后的输入神经元数据与权值数据之间的映射关系;

所述指令缓存单元 110 接收由所述控制器单元发送的目标指令;

所述指令处理单元 111 将所述目标指令译码为运算指令; 由所述运算单元 12 对所述输入神经元-权值对执行运算操作，得到运算结果。

进一步可选地，所述第一数据还包括稀疏参数; 还包括如下步骤:

稀疏单元 15 依据所述稀疏参数对所述运算结果进行稀疏处理，得到稀疏处理后的运算结果。

进一步可选地，所述稀疏参数包括稀疏模式；

所述映射单元 14 根据所述映射模式对所述输入神经元和所述权值进行映射处理，具体为：

当处于所述稀疏模式为稀疏模式 1 时，获取所述稀疏模式 1 对应的权值稀疏序列，并依据该权值稀疏序列对所述权值进行映射处理；

当处于所述稀疏模式为稀疏模式 2 时，获取所述稀疏模式 2 对应的神经元稀疏序列，并依据该神经元稀疏序列对所述输入神经元进行映射处理；

当处于所述稀疏模式为稀疏模式 3 时，获取所述稀疏模式 3 对应的权值稀疏序列和神经元稀疏序列，并依据该权值稀疏序列和神经元稀疏序列对所述输入神经元和所述权值数据进行映射处理。

进一步可选地，所述稀疏参数还包括稀疏率，所述稀疏单元 15 依据所述稀疏参数对所述运算结果进行稀疏处理，具体为：

对神经元数据的元素绝对值排序，根据稀疏率计算获得需要稀疏的元素个数，根据需要稀疏的元素个数对排序后的神经元数据的元素作稀疏处理，并将稀疏后的稀疏神经元数据和神经元稀疏序列发送至所述控制器单元 11。

进一步可选地，所述稀疏单元 15 依据所述稀疏参数对所述运算结果进行稀疏处理，具体为：

神经元数据为 0 的元素保持 0 不变，神经元数据在预设取值区间内的元素置为 0。

上述本申请所描述的运算装置和方法，支持神经网络的映射处理，可根据不同的实际运用情况，采用不同的映射模式，能够实现节省存储资源和运算资源的目的，另外，支持神经网络的稀疏处理及多种数据表示形式的神经网络，可根据不同的实际运用情况，采用不同的数据表示形式和稀疏处理，进一步节省在神经网络性能与运算和存储资源上提高性能，达到最优效果，对数据做稀疏处理，减小运算单元的运算负荷，加快运算速度。

可选地，图 1A 所描述的运算装置还可以用于实现如下向量排序功能，具体如下：

如所述输入神经元数据为向量；本申请实施例中，所述第一指令包括向量

排序指令，以及所述第一数据包括待排序数据向量以及待排序数据向量的中间结果。其中，第一指令可为特指要发射的指令，或者，指令队列中最前列的指令。

所述指令处理单元 111，用于将所述向量排序指令译码成所述运算单元 12 执行的微指令；

所述运算单元 12，还具体用于根据所述微指令将所述待排序数据向量或所述中间结果进行排序，得到与所述待排序数据向量等长度的排序后的向量。

可选地，运算单元 12 可以通过指令配置的方式动态选择排序方法完成向量排序运算。排序方法可以包括以下至少一种：冒泡排序、选择排序、快速排序、归并排序或者二分排序。

可选地，控制器单元 11 从存储单元 10 中获取所要执行的第一指令，若所述第一指令为向量排序指令，则根据向量排序类型、待排序数据向量的长度、向量排序指令的源操作数地址、目的操作数地址、向量长度和排序类型生成微指令，由运算单元 12 响应微指令执行排序操作。

如下表所示，示出了向量排序指令的一种可选地格式，具体如下：

操作数 OP	向量源地址 SRC	向量目的地址 DST	向量长度 LEN	排序类型

其中，操作码 OP，长度为 k 位，例如，具体内容为$b_1 b_2 \ldots b_k$，操作码 OP 用于指明指令所作的操作为向量排序操作，若某一指令的前 k 位与$b_1 b_2 \ldots b_k$不一致，则表明该指令用于实现其他运算操作，若某一指令的前 k 位于$b_1 b_2 \ldots b_k$一致，则表明该指令为向量排序指令。

向量源地址 SRC，长度为 q 位，当操作码为$b_1 b_2 \ldots b_k$时，表示待排序数据向量在存储单元中的地址，q 大于 0；

向量目的地址 DST，长度为 q 位，当操作码为$b_1 b_2 \ldots b_k$时，表示排序后的向量在存储单元中的地址，q 大于 0；

向量长度 LEN，长度为 r 位，当操作码为$b_1 b_2 \ldots b_k$时，表示待排序数据向量的长度，r 大于 0；

排序类型，长度为 1 位或多位，当操作码为$b_1 b_2 \ldots b_k$时，若排序类型的最高位为 0，表示由小到大进行排序。若排序类型的最高位为 1，表示由大到

小进行排序。若排序类型域的长度为 a(a>1)位，排序类型的低 a-1 位指明排序指令所采用的排序方法。

可选地，所述运算单元 12 根据所述微指令将所述待排序数据向量或所述中间结果进行排序具体为：

步骤 A：若排序后得到的为待排序数据向量的中间结果，则将所述待排序数据向量的中间结果写回到存储单元的源地址，并重复执行所述步骤 A，直到所述得到待排序数据向量的最后结果，则跳转到步骤 B；步骤 B：若排序得到的为待排序数据向量的最后结果，将所述待排序数据向量的最后结果根据所述向量排序指令提供的目的操作数地址写回到存储单元的数据 I/O 单元，操作结束。

可选地，所述运算单元 12 包括由 n 个向量归并单元，n 为大于等于 2 的整数；n 个向量归并单元每次从所述存储单元 10 中读取不大于 2n 个已经归并的子向量或者有序子向量，并进行归并，转存入所述存储单元中，直到已经归并的子向量的长度等于所述待排序数据向量长度，形成排序后的向量。

其中，如图 1F，图 1F 给出了运算单元 12 的具体细化结构，运算单元可包含 n 个向量归并单元，n 个向量归并单元从存储单元 10 中读取不大于 2n 个已经归并的子向量或者有序子向量，并进行归并，转存入存储单元 10 中，直到已经归并的子向量的长度等于所述待排序数据向量长度，形成排序后的向量。

如图 1G 所示，向量归并单元的具体结构可参见图 1G 所示的第 1 个向量归并单元，该向量归并单元包括控制信号 CTRL（连接控制器单元 11），输入向量 D_1 和 D_2（连接存储单元 10），输出数据为 OUT（用于连接存储单元 10 中的数据 I/O 单元）。其中，CTRL 用于设置向量归并单元的操作类型和输入向量 D_1 和 D_2 的长度 len_1 和 len_2。其中，操作类型可以用来描述进行归并的顺序。举例而言，操作类型可以包括 0 和 1，操作类型 0 可以用来表示向量归并单元根据向量由小到大的顺序进行归并，操作类型 1 可以用来表示向量归并单元根据向量由大到小的顺序进行归并。在其他实施例中，也可以用操作类型 1 来表示向量归并单元根据向量由小到大的顺序进行归并，操作类型 0 来表示向量归并单元根据向量由大到小的顺序进行归并。可以理解，操作类型

还可以根据具体需求进行设置。以两个输入向量为例，若其中一个输入向量的长度为 0 时，则可以直接输出另一个向量。

可选地，所述运算单元 12 执行步骤 A 具体为:

步骤 A1、初始化归并次数 i 为 1;

步骤 A2、由所述 n 个向量归并单元进行计算，在第 i 次归并所述待排序数据向量或所述中间结果时，从所述存储单元中获取所述待排序数据向量或所述中间结果，将所述待排序数据向量或中间结果按顺序分成$\lceil m/2^{i-1} \rceil$份，对向量进行两两归并，除最后一份外，每个向量长度为2^{i-1}；m 为待排序数据向量的长度;

步骤 A3、若归并次数$i < \lceil log_2 m \rceil$，则将归并次数加一，并将处理后的中间结果写回到存储单元源地址，重复执行步骤 A2-A3，直到$i = \lceil log_2 m \rceil$，则跳转到步骤 B；可以理解，当$i = \lceil log_2 m \rceil$时，则排序得到的为待排序数据向量的最后结果。

所述运算单元 12 执行步骤 B，具体为:

若归并次数$i = \lceil log_2 m \rceil$，若只存在分配后的两份待排序数据向量，则经所述 n 个向量归并单元中的第一个向量归并单元归并后，得到的向量为已排序向量，将排序后的结果根据所述向量排序指令提供的目的操作数地址写入到数据输出单元中，操作结束。

可选地，所述运算单元 12 对向量进行两两归并，具体为:

根据所述向量排序指令提供的源操作数地址按顺序编号 1、2、…、$\lceil m/2^{i-1} \rceil$，将编号为 2*j-1、2*j 的向量分配给第（(j -1) mod n ）+ 1 个向量归并单元进行处理，其中 j>0。

可选地，所述待排序数据向量为预处理阶段测试数据特征矩阵对应的特征值向量和分类结果的概率向量。

举例说明下，当向量 D_1 和 D_2 且操作类型为 0 分别为 2、4、6、7 和 3、3、8、9 时，归并的过程如下步骤 1-步骤 7，具体如下:

步骤 1:

D_1: 2 4 6 7 len_1:4

D_2: 3 3 8 9 len_2:4

输出向量: 2

步骤 2:

D_1: 4 6 7 len_1:3

D_2: 3 3 8 9 len_2:4

输出向量: 2 3

步骤 3:

D_1: 4 6 7 len_1:3

D_2: 3 8 9 len_2:3

输出向量: 2 3 3

步骤 4:

D_1: 4 6 7 len_1:3

D_2: 8 9 len_2:2

输出向量: 2 3 3 4

步骤 5:

D_1: 6 7 len_1:2

D_2: 8 9 len_2:2

输出向量: 2 3 3 4 6

步骤 6:

D_1: 7 len_1:1

D_2: 8 9 len_2:2

输出向量: 2 3 3 4 6 7

步骤 7:

D_1: len_1:0

D_2: 8 9 len_2:2

输出向量: 2 3 3 4 6 7 8 9

向量归并结束。

再举例说明下，对于向量排序指令(sort_op, src, dst, 9, 0)，假设数据存储单元 1 地址 src 开始，连续存放的 9 个数据分别为 9、1、5、3、4、2、6、8、7，向量归并单元数量为 2。根据排序类型的最高位，进行由小到大的排序，根据排序类型的低 m-1 位全 0，进行排序的类型为归并排序。在运算过程，每次每个向量归并单元对应输入的向量、地址以及输出地址如下所示:

第一次合并:

向量合并单元 1	向量 1	9	4	7
	向量 1 地址	src	src +4	src +8
	向量 2	1	2	
	向量 2 地址	src+1	src + 5	src + 9
	输出向量	1 9	2 4	7
	输出向量地址	src	src + 4	src + 8
向量合并单元 2	向量 1	5	6	
	向量 1 地址	src + 2	src + 6	
	向量 2	3	8	
	向量 2 地址	src + 3	src + 7	
	输出向量	3 5	6 8	
	输出向量地址	src +3	src + 6	

合并后的向量为 1 9 3 5 2 4 6 8 7;

第二次合并:

向量合并单元 1	向量 1	1 9	7
	向量 1 地址	src	src +8
	向量 2	3 5	
	向量 2 地址	src+2	src + 10
	输出向量	1 3 5 9	7
	输出向量地址	src	src + 8
向量合并单元 2	向量 1	2 4	
	向量 1 地址	src + 4	
	向量 2	6 8	
	向量 2 地址	src + 6	
	输出向量	2 4 6 8	
	输出向量地址	src + 4	

合并后的向量为 1 3 5 9 2 4 6 8 7;

第三次合并:

向量合并单元 1	向量 1	1 3 5 9
	向量 1 地址	src
	向量 2	2 4 6 8
	向量 2 地址	src+4
	输出向量	1 2 3 4 5 6 8 9
	输出向量地址	src
向量合并单元 2	向量 1	7
	向量 1 地址	src + 8
	向量 2	
	向量 2 地址	src + 12
	输出向量	7
	输出向量地址	src + 8

合并后的向量为： 1 2 3 4 5 6 7 8 7

第四次合并:

向量合并单元 1	向量 1	1 2 3 4 5 6 7 8 7
	向量 1 地址	src
	向量 2	7
	向量 2 地址	src+8
	输出向量	1 2 3 4 5 6 7 7 8
	输出向量地址	dst

合并后的向量为：1 2 3 4 5 6 7 7 8，并将其送入到数据 I/O 单元 13 中。

可选地，基于上述运算装置，可以实现如下运算方法，具体如下

所述指令处理单元 111 将所述向量排序指令译码成所述运算单元执行的微指令;

所述运算单元 12 根据所述微指令将所述待排序数据向量或所述中间结果进行排序，得到与所述待排序数据向量等长度的排序后的向量。

可选地，所述运算单元 12 根据所述微指令将所述待排序数据向量或所述中间结果进行排序具体为:

步骤 A: 若排序后得到的为待排序数据向量的中间结果，则将所述待排序数据向量的中间结果写回到存储单元的源地址，并重复执行所述步骤 A，直到所述得到待排序数据向量的最后结果，则跳转到步骤 B; 步骤 B: 若排序得到的为待排序数据向量的最后结果，将所述待排序数据向量的最后结果根据所述向量排序指令提供的目的操作数地址写回到存储单元的数据 I/O 单元，操作结束。

进一步可选地，所述运算单元 12 包括 n 个向量归并单元构成，其中，n 为大于等于 2 的整数，所述 n 个向量归并单元用于从所述存储单元中读取不大于 2n 个已经归并的子向量或者有序子向量，并进行归并，将归并后的结果转存入所述存储单元中，直到已经归并的子向量的长度等于所述待排序数据向量长度，形成排序后的向量。

进一步可选地，所述运算单元 12 执行步骤 A 具体为:

步骤 A1、初始化归并次数 i 为 1;

步骤 A2、由所述 n 个向量归并单元进行计算，在第 i 次归并所述待排序数据向量或所述中间结果时，从所述存储单元中获取所述待排序数据向量或所述中间结果，将所述待排序数据向量或中间结果按顺序分成$\lceil m/2^{i-1} \rceil$份，对向量进行两两归并，除最后一份外，每个向量长度为2^{i-1}；m 为待排序数据向量的长度;

步骤 A3、若归并次数$i < \lceil log_2 m \rceil$，则将归并次数加一，并将处理后的中间结果写回到存储单元源地址，重复执行步骤 A2-A3，直到$i = \lceil log_2 m \rceil$，则跳转到步骤 B;

所述运算单元 12 执行步骤 B，具体为:

若归并次数$i = \lceil log_2 m \rceil$，若只存在分配后的两份待排序数据向量，则经所述 n 个向量归并单元中的第一个向量归并单元归并后，得到的向量为已排序向量，将排序后的结果根据所述向量排序指令提供的目的操作数地址写入到数据输出单元中，操作结束。

进一步可选地，所述运算单元 12 对向量进行两两归并，具体为:

根据所述向量排序指令提供的源操作数地址按顺序编号 1、2、...、$\lceil m/2^{i-1} \rceil$，将编号为 2*j-1、2*j 的向量分配给第（（j -1）mod n ）+ 1 个向量归并单元进行处理，其中 j>0。

进一步可选地，所述待排序数据向量为预处理阶段测试数据特征矩阵对应的特征值向量和分类结果的概率向量。

进一步可选地，所述第一指令包括下述指令中的一个或任意组合: 向量间与指令 VAV、向量内与指令 VAND、向量间或指令 VOV、向量内或指令 VOR、向量指数指令 VE、向量对数指令 VL、向量大于判定指令 VGT、向量等于判定指令 VEQ、向量非指令 VINV、向量选择合并指令 VMER、向量最大值指令 VMAX、标量扩展指令 STV、标量替换向量指令 STVPN、向量替换标量指令 VPNTS、向量检索指令 VR、向量点积指令 VP、随机向量指令 RV、循环移位指令 VCS、向量加载指令 VLOAD、向量存储指令 VS、向量搬运指令 VMOVE、矩阵检索指令 MR、矩阵加载指令 ML、矩阵存储指令 MS、矩阵搬

运指令 MMOVE。

进一步可选地，所述装置用于稀疏神经网络运算或者稠密神经网络运算。

采用本申请实施例，将向量排序指令译码成运算单元执行的微指令，根据微指令将待排序数据向量或所述中间结果进行排序，得到与待排序数据向量等长度的排序后的向量。相对于现有技术中，相关操作串行执行，很难利用排序算法的可并行性，运算速度较慢，且向量排序算法会被分成译码成一系列的指令序列，译码的开销也很大，本申请可以并行执行排序，且降低译码开销，提升了排序效率。

在一种可选实施例中，图 1H 作为图 1A 所示的运算装置的一种变型结构，其运算单元 12 如图 1H 所示，可以包括分支处理电路 1003；其具体的连接结构如图 1I 所示，其中，

主处理电路 1001 与分支处理电路 1003 连接，分支处理电路 1003 与多个从处理电路 1002 连接；

分支处理电路 1003，用于执行转发主处理电路 1001 与从处理电路 1002 之间的数据或指令。

在另一种可选实施例中，运算单元 12 如图 1C 所示，可以包括一个主处理电路 1001 和多个从处理电路 1002。在一个实施例里，如图 1C 所示，多个从处理电路呈阵列分布；每个从处理电路与相邻的其他从处理电路连接，主处理电路连接所述多个从处理电路中的 k 个从处理电路，所述 k 个基础电路为：第 1 行的 n 个从处理电路、第 m 行的 n 个从处理电路以及第 1 列的 m 个从处理电路。

K 个从处理电路，用于在所述主处理电路以及多个从处理电路之间的数据以及指令的转发。

可选的，如图 1J 所示，该主处理电路还可以包括：转换处理电路 1010、激活处理电路 1011、加法处理电路 1012 中的一种或任意组合；

转换处理电路 1010，用于将主处理电路接收的数据块或中间结果执行第一数据结构与第二数据结构之间的互换（例如连续数据与离散数据的转换）；或将主处理电路接收的数据块或中间结果执行第一数据类型与第二数据类型

之间的互换（例如定点类型与浮点类型的转换）;

激活处理电路 1011，用于执行主处理电路内数据的激活运算;

加法处理电路 1012，用于执行加法运算或累加运算。

所述主处理电路，用于将将一个输入数据分配成多个数据块，将所述多个数据块中的至少一个数据块以及多个运算指令中的至少一个运算指令发送给所述从处理电路;

所述多个从处理电路，用于依据该运算指令对接收到的数据块执行运算得到中间结果，并将运算结果传输给所述主处理电路;

所述主处理电路，用于将多个从处理电路发送的中间结果进行处理得到该运算指令的结果，将该运算指令的结果发送给所述控制器单元。

所述从处理电路包括：乘法处理电路;

所述乘法处理电路，用于对接收到的数据块执行乘积运算得到乘积结果;

转发处理电路（可选地），用于将接收到的数据块或乘积结果转发。

累加处理电路，所述累加处理电路，用于对该乘积结果执行累加运算得到该中间结果。

另一个实施例里，该运算指令为矩阵乘以矩阵的指令、累加指令、激活指令等等计算指令。

本申请提供的计算装置设置了互联单元，此互联单元能够根据运算指令的需要将运算单元内的计算器组合连接得到与该运算指令对应的计算拓扑结构，进而在后续的运算单元运算时无需对计算的中间数据执行存储或提取操作，此结构实现单一指令即能实现一次输入即能够进行多次计算器的运算得到计算结果的优点，提高了计算效率。

其中，数据转换单元 16 从装置外获取结构图中的部分节点，通过控制器单元 11 判断节点是否经过运算单元 12 处理，如果已经处理，将该节点舍弃，不做任何操作；如果没有处理，对该节点进行节点格式的转换，转换完成后，将其写入到存储单元 10 中。控制器单元 11 将指令从运算装置外部读入，不经转换，写入到存储单元 10 中。第一次从装置外获取的结构图（如图 1K 所示，图 1K 给出了一种结构图的示意图）中节点为源节点 s，第 i 次获取的点为第

i-1（i>1）次计算后得到的候选节点的邻接节点且此邻接节点未被运算单元处理。可选地，它将从装置外获取的结构图的节点 n 转化为如下的格式:

$$(\mathrm{Addr}(\mathrm{before}(n)),F(n),n,vis)$$

其中，before(n)表示节点 n 的前驱节点，$\mathrm{Addr}(\mathrm{before}(n))$表示节点 n 的前驱节点在存储单元 10 中的地址，F(n)表示从源节点 s 到节点 n 的路径上产生的总代价，n 表示节点的属性，用于计算单个节点 n 所产生的代价，vis 表示此节点是否被访问过，例如，未被访问过记作 0，被访问过记作 1，对于源节点before(n)为 n 本身，F（n）和 vis 都设置为 0。

存储单元 10，用于从数据转换单元数据转换单元 14 中获取指令和转换后的数据，为运算单元 12 提供数据，存储经由运算单元 12 处理后的数据，得到近似最优路径结果并存储，最后将近似最优路径结果写回到装置外部。

运算单元 12，从存储单元 10 中获取节点信息中 vis=0 的节点，即未被访问过的节点，将此节点的前驱节点的部分信息整合到此节点构成如下的格式:

$$(F(\mathrm{before}(n)),F(n),n,vis)$$

其中，$F(\mathrm{before}(n))$表示从源节点 s 到 n 的前驱节点的路径对应的代价值，$F(n)=0$。在运算单元中，基于预设代价函数计算节点 n 所产生的代价值f(n)，然后，得到源节点到节点 n 的路径对应的总代价值 $F(n)=f(n)+F(\mathrm{before}(n))$. 此时送入运算单元 12 的节点有 m 个，分别表示为$n_1,n_2,\ldots n_m$,可计算得到 m 个路径对应的代价值$F(n_1),F(n_2),\ldots F(n_m)$。将对应的 m 个节点按照代价值$F(n_1),F(n_2),\ldots F(n_m)$从小到大的顺序进行排序得到$n_1',n_2',\ldots n_m'$。判断源节点 s 到$n_1'$的路径是否构成完整的近似最优路径，如果构成，则对控制器单元 11 发送运算终止指令，并将n_1'对应的节点信息$(\mathrm{Addr}(\mathrm{before}(n)),F(n),n,1)$传送到存储单元 10 中。本申请实施例中的预设代价函数可以为以下至少一种:均方误差代价函数、交叉熵代价函数，或者，神经网络中的代价函数。

可选的，假设运算装置允许最大候选节点数为 k。当$m \le k$时，则可以将对应的 m 个节点都作为候选节点，将更新后的$(\mathrm{Addr}(\mathrm{before}(n)),F(n),n,1)$写

入存储单元 10 中；当 $m > k$ 时，则可以将 $n_1', n_2', \ldots n_k'$ 对应的节点信息 $(Addr(before(n)), F(n), n, 1)$ 写回到存储单元 10 中。

运算单元 12 可以在本单元内部维持一个空的堆栈，在收到控制器单元 11 发送的整合指令之后，对节点信息 $(Addr(before(n)), F(n), n, 1)$ 进行整合，具体地，将节点 n 压入堆栈中，然后，从存储单元 10 中获取该堆栈顶部节点的前驱节点，并压入该堆栈，直到栈顶节点的信息中 before(n)为 n，即栈顶节点为图的源节点。然后，将堆栈中节点不断出栈，按照顺序送入到存储单元 10 中，存储单元 10 中获取的节点序列即为最终得到的近似最优路径。

控制器单元 11 通过存储单元 10 获取运算所需的指令，存储单元 10 读取上一次节点从运算装置外部存入的节点的尚未被运算单元 12 运算的节点，其控制运算单元 12 进行数据运算，并接收运算单元 12 发送的运算终止指令，控制运算单元 12 与存储单元 10 之间的数据传输。

请参见图 1L，图 1L 中运算单元 12 包括节点分发单元 41、代价函数计算单元 42、排序单元 43 和终止判断单元 44。

其中，节点分发单元 41 将存储单元 10 获取的节点 $n_1, n_2, \ldots n_m$ 分别分配给 L 个代价函数计算单元 42，并由其计算对应的路径代价，其中，前 L-1 个代价函数计算单元分别分配 $\lceil m/L \rceil$ 个节点，第 L 个代价函数计算单元分配 $m - \lceil m/L \rceil$ 个节点，其中“⌈ ⌉”表示向上取整。

如图 1L 中，图中共计 L 个代价函数计算单元，每个代价函数计算单元 42 可以实现独立实现从源节点到对应路径的代价值。每个代价函数计算单元 42 对由节点分发单元 41 分配的节点，计算得到对应的节点号-路径代价值对 $(n_1, F(n_1)), (n_2, F(n_2)), \ldots (n_m, F(n_m))$，计算路径代价的函数根据实际需要，由控制器单元 11 进行设置，然后，将计算得到的节点号-路径代价值对 $(n_1, F(n_1)), (n_2, F(n_2)), \ldots (n_m, F(n_m))$ 传送到排序单元 43 中。

可选地，代价函数计算单元可以包括乘法器以及加法器。

排序单元 43 将从代价函数计算单元 32 获取的各节点的节点号-路径代价值对 $(n_1, F(n_1)), (n_2, F(n_2)), \ldots (n_m, F(n_m))$，根据路径代价值从小到大进行排

序，得到排序后的节点号-路径代价值对$(n_1',F(n_1))',(n_2',F(n_2))',\ldots(n_m',F(n_m)')$，并将传送到终止判断单元34。

终止判断单元44从排序单元43中获取排序后的节点号-路径代价值对，根据$(n_1',F(n_1))'$判断是否已经取得近似最优路径，其中，判断是否已经取得近似最优路径的方法根据实际问题由控制器单元11设置。如果已经取得近似最优路径，则终止判断单元34对控制器单元11发送运算终止指令，并将n_1'对应的节点信息$(\mathrm{Addr}(\mathrm{before(n)}),\mathrm{F(n)},\mathrm{n},1)$传送到结果存储单元10中。否则，假设装置允许的最大候选节点数为k，若$\mathrm{m} \le \mathrm{k}$,则将对应的m个节点都作为候选节点将更新后的$(\mathrm{Addr}(\mathrm{before(n)}),\mathrm{F(n)},\mathrm{n},1)$写回到存储单元10中，若$\mathrm{m} > k$,则将$n_1',n_2',\ldots n_\mathrm{k}'$对应的节点信息$(\mathrm{Addr}(\mathrm{before(n)}),\mathrm{F(n)},\mathrm{n},1)$写回到存储单元10中。

举例说明下，如下提供了一种最优路径寻找方法，具体如下:

步骤1，从运算装置外部获取运算所需的运算指令，经由数据转换单元16存储到存储单元10中，传输到控制器单元11。

步骤2，从运算装置外部将原始图中部分节点传送到数据转换单元16中，装置外部的结构图可以是邻接表、邻接矩阵、顶点对或者其他形式。其中，第一次传送时只传送源节点s，之后传送时传输到数据转换单元1中节点为上一次经运算单元12筛选得到的候选节点的尚未被运算单元12处理过的邻接节点。在此过程中，由控制器单元11判断对应的节点是否已经被运算单元12处理过。数据转换单元16将传入的节点按照$(\mathrm{Addr}(\mathrm{before(n)}),\mathrm{F(n)},\mathrm{n},\mathrm{vis})$的格式进行转换，然后送至存储单元10中。

步骤3中，控制器单元11控制运算单元12从存储单元10中获取由数据转换单元16传入的尚未被处理的节点$n_1,n_2,\ldots n_m$，将各个节点与前驱节点的信息进行整合得到格式为$(\mathrm{F}(\mathrm{before(n)}),\mathrm{F(n)},\mathrm{n},\mathrm{vis})$的节点。然后，运算单元12基于预设代价函数计算节点n所产生的代价值f(n)，得到源节点到节点n的路径对应的总代价值$\mathrm{F(n)} = \mathrm{f(n)} + \mathrm{F(before(n))}$。分别计算m个路径对应的代价值$\mathrm{F}(n_1),F(n_2),\ldots F(n_m)$。将对应的m个节点按照代价值$\mathrm{F}(n_1),F(n_2),\ldots F(n_m)$从小到大的顺序进行排序得到$n_1',n_2',\ldots n_m'$。判断源节

点 s 到n_1'的路径是否构成完整的近似最优路径，如果构成，则对控制器单元 11 发送运算终止指令，并将n_1'对应的节点信息$(\mathrm{Addr}(\mathrm{before(n)}),\mathrm{F(n)},\mathrm{n},1)$传送到存储单元 10 中，转入步骤 4。否则，假设装置允许的最大候选节点数为 K，若 $\mathrm{m} \leq \mathrm{k}$，则将对应的 m 个节点都作为候选节点将更新后的$(\mathrm{Addr}(\mathrm{before(n)}),\mathrm{F(n)},\mathrm{n},1)$写回到存储单元 10 中，若$\mathrm{m} > k$，则将$n_1', n_2', \ldots n_\mathrm{k}'$对应的节点信息$(\mathrm{Addr}(\mathrm{before(n)}),\mathrm{F(n)},\mathrm{n},1)$写回到存储单元 10 中，转入到步骤 2 中。

步骤 4 中，控制器单元 11 在收到来自运算单元 12 的运算终止指令后，判断结果存储单元 10 是否已经从运算单元 12 中获取节点信息，如果没有取得节点信息，则一直循环判断，直到取得为止。结果存储单元 11 在单元内部维持一个空的堆栈，在取得运算结果后，将收到的来自运算单元 12 的节点信息$(\mathrm{Addr}(\mathrm{before(n)}),\mathrm{F(n)},\mathrm{n},1)$压入堆栈中。然后从存储单元 10 中获取堆栈顶部节点的前驱节点，并压入堆栈，重复此过程，直到栈顶节点的信息中 before(n)为 n，即栈顶节点为图的源节点。然后，存储单元 10 将堆栈中节点不断出栈，按照顺序送入到存储单元 10 中，存储单元 10 中获取的节点序列即为最终得到的近似最优路径。

在步骤 5 中，存储单元 10 将在控制器单元 11 的控制下，从存储单元 10 中获取的近似最优路径，并将其传输到装置外部。

采用本申请实施例，通过对结构图进行搜索，找到一条能够满足条件的近似最优路径，可以有效地减少空间消耗，并提高时间效率，且在计算路径的代价的过程中，采用多个代价函数计算单元同时进行计算，可以提高运算的并行性。

需要说明的是，上述运算装置不仅可以进行稀疏神经网络运算，还可以进行稠密神经网络运算。上述运算装置特别适用于稀疏神经网络的运算，是因为稀疏神经网络里 0 值数据或者绝对值很小的数据非常多。通过映射单元可以提出这些数据，在保证运算精度的情况下，可提高运算的效率。

需要指出的是，本申请实施例中提到的输入神经元和运算结果（或者输出神经元）并非是指整个神经网络的输入层中的神经元和输出层中的神经元，而

是对于神经网络中任意相邻的两层神经元，处于网络前馈运算下层中的神经元即为输入神经元，处于网络前馈运算上层中的神经元即为运算结果。以卷积神经网络为例，假设一个卷积神经网络有 L 层，K=1,2,3…L-1，对于第 K 层和第 K+1 层来说，第 K 层被称为输入层，该层中的神经元为上述输入神经元，第 K+1 层被称为输入层，该层中的神经元为上述运算结果，即除了顶层之外，每一层都可以作为输入层，其下一层为对应的输出层。

上述各单元可以是硬件电路包括数字电路，模拟电路等等。硬件电路的物理实现包括但不局限于物理器件，物理器件包括但不局限于晶体管，忆阻器等等。上述神经网络运算模块中的运算单元可以是任何适当的硬件处理器，比如 CPU、GPU、FPGA、DSP 和 ASIC 等等。上述存储单元、指令缓存单元，第一输入缓存单元、第二输入缓存单元和输出缓存单元均可以是任何适当的磁存储介质或者磁光存储介质，比如 RRAM, DRAM, SRAM, EDRAM, HBM, HMC 等等。

在一种可行的实施例中，本申请实施例提供了一种神经网络计算装置，该神经网络计算装置包括一个或多个如上述所示实施例所述的神经网络运算模块，用于从其他处理装置中获取待运算数据和控制信息，并执行指定的神经网络运算，将执行结果通过 I/O 接口传递给其他处理装置；

当所述神经网络计算装置包含多个所述神经网络运算模块时，所述多个所述神经网络运算模块间可以通过特定的结构进行连接并传输数据；

其中，多个所述运算装置通过 PCIE 总线进行互联并传输数据，以支持更大规模的神经网络的运算；多个所述运算装置共享同一控制系统或拥有各自的控制系统；多个所述运算装置共享内存或者拥有各自的内存；多个所述运算装置的互联方式是任意互联拓扑。

该神经网络计算装置具有较高的兼容性，可通过 PCIE 接口与各种类型的服务器相连接。

在一种可行的实施例中，本申请实施例提供了一种组合处理装置，该组合装置包括如上述神经网络计算装置，通用互联接口和其他处理装置。

上述神经网络计算装置与上述其他处理装置进行交互，共同完成用户指定的操作。参见图 2A，图 2A 为本申请实施例提供的一种组合处理装置的结构

示意图。如图 2A 所示，该组合处理装置包括上述神经网络计算装置 1601、通用互联接口 1602 和其他处理装置 1603。

其中，上述其他处理装置 1603 包括中央处理器（Central Processing Unit）、图形处理器（Graphics Processing Unit，GPU）、神经网络处理器等通用／专用处理器中的一种或以上的处理器类型。其他处理装置 1603 所包括的处理器数量不做限制。其他处理装置 1603 作为神经网络计算装置 1601 与外部数据和控制的接口，包括数据搬运，完成对本神经网络计算装置的开启、停止等基本控制；其他处理装置 1603 也可以和神经网络计算装置 1601 协作共同完成运算任务。

上述通用互联接口 1602，用于在所述神经网络计算装置 1601 与其他处理装置 1603 间传输数据和控制指令。该神经网络计算装置 1601 从其他处理装置 1603 中获取所需的输入数据，写入神经网络计算装置 1601 片上的存储装置；可以从其他处理装置 1603 中获取控制指令，写入神经网络计算装置 1601 片上的控制缓存；也可以读取神经网络计算装置 1601 的存储装置中的数据并传输给其他处理装置 1603。

可选的，如图 2B 所示，上述组合处理装置还包括存储装置 1604，用于保存在本运算单元／运算装置或其他运算单元所需要的数据，尤其适用于所需要运算的数据在本神经网络计算装置 1601 或其他处理装置 1603 的内部存储中无法全部保存的数据。

上述组合装置可以作为手机、机器人、无人机等智能设备的片上系统，有效降低控制部分的核心面积，提高处理速度，降低整体功耗。

在一种可行的实施例中，本申请实施例提供了一种神经网络芯片，该神经网络芯片包括任一所示实施例所述的运算装置，或者上述神经网络计算装置或者上述组合处理装置。

在一种可行的实施例中，本申请实施例提供了一种神经网络芯片封装结构，该神经网络芯片封装结构包括上述神经网络芯片。

在一种可行的实施例中，本申请实施例提供了一种板卡，该板卡包括上述神经网络芯片封装结构。

在一种可行的实施例中，本申请实施例提供了一种电子装置，该电子装置

包括上述板卡。

其中，该电子装置包括：数据处理装置、机器人、电脑、打印机、扫描仪、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、可穿戴设备交通工具、家用电器、和/或医疗设备。

上述交通工具包括飞机、轮船和/或车辆；上述家用电器包括电视、空调、微波炉、冰箱、电饭煲、加湿器、洗衣机、电灯、燃气灶、油烟机；所述医疗设备包括核磁共振仪、B 超仪和/或心电图仪。

本申请实施例还提供一种计算机存储介质，其中，该计算机存储介质可存储有程序，该程序执行时包括上述方法实施例中记载的任何一种神经网络运算方法的部分或全部步骤。需要说明的是，对于前述的各方法实施例，为了简单描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本申请并不受所描述的动作顺序的限制，因为依据本申请，某些步骤可以采用其他顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述的实施例均属于可选 实施例，所涉及的动作和模块并不一定是本申请所必须的。

在上述实施例中，对各个实施例的描述都各有侧重，某个实施例中没有详述的部分，可以参见其他实施例的相关描述。

在本申请所提供的几个实施例中，应该理解到，所揭露的装置，可通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，装置或单元的间接耦合或通信连接，可以是电性或其它的形式。

所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

另外，在本申请各个实施例中的各功能单元可以集成在一个处理单元中，

也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用软件程序模块的形式实现。

所述集成的单元如果以软件程序模块的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储器中。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的全部或部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储器中，包括若干指令用以使得一台计算机设备（可为个人计算机、服务器或者网络设备等）执行本申请各个实施例所述方法的全部或部分步骤。而前述的存储器包括：U盘、只读存储器（ROM，Read-Only Memory）、随机存取存储器（RAM，Random Access Memory）、移动硬盘、磁碟或者光盘等各种可以存储程序代码的介质。

本领域普通技术人员可以理解上述实施例的各种方法中的全部或部分步骤是可以通过程序来指令相关的硬件来完成，该程序可以存储于一计算机可读存储器中，存储器可以包括：闪存盘、只读存储器（英文：Read-Only Memory，简称：ROM）、随机存取器（英文：Random Access Memory，简称：RAM）、磁盘或光盘等。

以上对本申请实施例进行了详细介绍，本文中应用了具体个例对本申请的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本申请的方法及其核心思想；同时，对于本领域的一般技术人员，依据本申请的思想，在具体实施方式及应用范围上均会有改变之处，综上所述，本说明书内容不应理解为对本申请的限制。

权 利 要 求

1、一种运算装置，其特征在于，所述运算装置包括存储单元、运算单元和控制器单元，其中，

所述存储单元，用于存储数据和指令;

所述控制器单元，用于从所述存储单元中提取第一指令以及所述第一指令对应的第一数据，所述第一数据包括输入神经元数据和权值数据，所述第一指令包括排序指令或者稀疏处理指令;

所述运算单元，用于响应所述第一指令，对所述输入神经元数据和所述权值数据执行所述第一指令对应的运算操作，得到运算结果。

2、根据权利要求 1 所述的装置，其特征在于，所述控制器单元包括：指令缓存单元和指令处理单元，所述指令缓存单元用于对指令进行缓存，所述指令处理单元用于实现译码功能。

3、根据权利要求 2 所述的装置，其特征在于，所述装置还包括：配置解析单元和映射单元，在所述第一指令为所述稀疏处理指令以及所述第一数据还包括预设配置数据时，其中，

所述配置解析单元，用于根据所述预设配置数据设置映射模式;

所述映射单元，用于根据所述映射模式对所述输入神经元和所述权值数据进行映射处理，得到输入神经元-权值对，所述输入神经元-权值对为映射处理后的输入神经元数据与权值数据之间的映射关系;

所述指令缓存单元，用于接收由所述控制器单元发送的目标指令;

所述指令处理单元，用于将所述目标指令译码为运算指令；由所述运算单元对所述输入神经元-权值对执行运算操作，得到运算结果。

4、根据权利要求 1 所述的装置，其特征在于，所述第一数据还包括稀疏参数；所述装置还包括:

稀疏单元，用于依据所述稀疏参数对所述运算结果进行稀疏处理，得到稀疏处理后的运算结果。

5、根据权利要求 4 所述的装置，其特征在于，所述稀疏参数包括稀疏模式;

所述映射单元根据所述映射模式对所述输入神经元和所述权值进行映射处理，具体为：

当处于所述稀疏模式为稀疏模式 1 时，获取所述稀疏模式 1 对应的权值稀疏序列，并依据该权值稀疏序列对所述权值进行映射处理；

当处于所述稀疏模式为稀疏模式 2 时，获取所述稀疏模式 2 对应的神经元稀疏序列，并依据该神经元稀疏序列对所述输入神经元进行映射处理；

当处于所述稀疏模式为稀疏模式 3 时，获取所述稀疏模式 3 对应的权值稀疏序列和神经元稀疏序列，并依据该权值稀疏序列和神经元稀疏序列对所述输入神经元和所述权值数据进行映射处理。

6、根据权利要求 4 或 5 所述的装置，其特征在于，所述稀疏参数还包括稀疏率，所述稀疏单元依据所述稀疏参数对所述运算结果进行稀疏处理，具体为：

对神经元数据的元素绝对值排序，根据稀疏率计算获得需要稀疏的元素个数，根据需要稀疏的元素个数对排序后的神经元数据的元素作稀疏处理，并将稀疏后的稀疏神经元数据和神经元稀疏序列发送至所述控制器单元。

7、根据权利要求 4 或 5 所述的装置，其特征在于，所述稀疏单元依据所述稀疏参数对所述运算结果进行稀疏处理，具体为：

神经元数据为 0 的元素保持 0 不变，神经元数据在预设取值区间内的元素置为 0。

8、根据权利要求 2 所述的装置，其特征在于，如所述输入神经元数据为向量；所述第一指令为向量排序指令，以及所述第一数据为待排序数据向量以及待排序向量的中间结果；所述控制器单元包括指令处理单元；

所述指令处理单元，用于将所述向量排序指令译码成所述运算单元执行的微指令；

所述运算单元，还具体用于根据所述微指令将所述待排序数据向量或所述中间结果进行排序，得到与所述待排序数据向量等长度的排序后的向量。

9、根据权利要求 8 所述的装置，其特征在于，所述运算单元根据所述微指令将所述待排序数据向量或所述中间结果进行排序具体为：

步骤 A: 若排序后得到的为待排序数据向量的中间结果，则将所述待排序数据向量的中间结果写回到存储单元的源地址，并重复执行所述步骤 A，直到所述得到待排序数据向量的最后结果，则跳转到步骤 B；步骤 B: 若排序得到的为待排序数据向量的最后结果，将所述待排序数据向量的最后结果根据所述向量排序指令提供的目的操作数地址写回到存储单元的数据 I/O 单元，操作结束。

10、根据权利要求 9 所述的装置，其特征在于，所述运算单元包括 n 个向量归并单元构成，其中，n 为大于等于 2 的整数，所述 n 个向量归并单元用于从所述存储单元中读取不大于 2n 个已经归并的子向量或者有序子向量，并进行归并，将归并后的结果转存入所述存储单元中，直到已经归并的子向量的长度等于所述待排序数据向量长度，形成排序后的向量。

11、根据权利要求 10 所述的装置，其特征在于，所述运算单元执行步骤 A 具体为:

步骤 A1、初始化归并次数 i 为 1;

步骤 A2、由所述 n 个向量归并单元进行计算，在第 i 次归并所述待排序数据向量或所述中间结果时，从所述存储单元中获取所述待排序数据向量或所述中间结果，将所述待排序数据向量或中间结果按顺序分成$\lceil m/2^{i-1} \rceil$份，对向量进行两两归并，除最后一份外，每个向量长度为2^{i-1}；m 为待排序数据向量的长度;

步骤 A3、若归并次数$i < \lceil log_2 m \rceil$，则将归并次数加一，并将处理后的中间结果写回到存储单元源地址，重复执行步骤 A2-A3，直到$i = \lceil log_2 m \rceil$，则跳转到步骤 B;

所述运算单元执行步骤 B，具体为:

若归并次数$i = \lceil log_2 m \rceil$，若只存在分配后的两份待排序数据向量，则经所述 n 个向量归并单元中的第一个向量归并单元归并后，得到的向量为已排序向量，将排序后的结果根据所述向量排序指令提供的目的操作数地址写入到数据输出单元中，操作结束。

12、根据权利要求 11 所述的装置，其特征在于，所述运算单元对向量进行两两归并，具体为：

根据所述向量排序指令提供的源操作数地址按顺序编号 1、2、…、$\lceil m/2^{i-1} \rceil$，将编号为 2*j-1、2*j 的向量分配给第（（j -1）mod n ）+ 1 个向量归并单元进行处理，其中 j>0。

13、根据权利要求 6-12 任一项所述的装置，其特征在于，所述待排序数据向量为预处理阶段测试数据特征矩阵对应的特征值向量和分类结果的概率向量。

14、根据权利要求 1-13 所述的装置，其特征在于，所述第一指令包括下述指令中的一个或任意组合：向量间与指令 VAV、向量内与指令 VAND、向量间或指令 VOV、向量内或指令 VOR、向量指数指令 VE、向量对数指令 VL、向量大于判定指令 VGT、向量等于判定指令 VEQ、向量非指令 VINV、向量选择合并指令 VMER、向量最大值指令 VMAX、标量扩展指令 STV、标量替换向量指令 STVPN、向量替换标量指令 VPNTS、向量检索指令 VR、向量点积指令 VP、随机向量指令 RV、循环移位指令 VCS、向量加载指令 VLOAD、向量存储指令 VS、向量搬运指令 VMOVE、矩阵检索指令 MR、矩阵加载指令 ML、矩阵存储指令 MS、矩阵搬运指令 MMOVE。

15、根据权利要求 1-14 任一项所述的装置，其特征在于，所述装置用于稀疏神经网络运算或者稠密神经网络运算。

16、一种神经网络计算装置，其特征在于，所述神经网络计算装置包括一个或多个如权利要求 1-15 任一项所述的运算装置，用于从其他处理装置中获取待运算数据和控制信息，并执行指定的神经网络运算，将执行结果通过 I/O 接口传递给其他处理装置；

当所述神经网络计算装置包含多个所述运算装置时，所述多个所述运算装置间可以通过特定的结构进行连接并传输数据；

其中，多个所述运算装置通过快速外部设备互连总线 PCIE 总线进行互联并传输数据，以支持更大规模的神经网络的运算；多个所述运算装置共享同一控制系统或拥有各自的控制系统；多个所述运算装置共享内存或者拥有各自的内存；多个所述运算装置的互联方式是任意互联拓扑。

17、一种组合处理装置，其特征在于，所述组合处理装置包括如权利要求16所述的神经网络计算装置，通用互联接口和其他处理装置;

所述神经网络计算装置与所述其他处理装置进行交互，共同完成用户指定的操作。

18、一种神经网络芯片，其特征在于，所述神经网络芯片包括如权利要求16所述的神经网络计算装置或如权利要求17所述的组合处理装置。

19、一种板卡，其特征在于，所述板卡包括如权利要求18所述的神经网络芯片。

20、一种电子装置，其特征在于，所述电子装置包括如权利要求18所述的神经网络芯片或者如权利要求19所述的板卡。

21、一种运算方法，其特征在于，应用于运算装置，所述运算装置包括存储单元、运算单元和控制器单元，其中，

所述存储单元存储数据和指令;

所述控制器单元从所述存储单元中提取第一指令以及所述第一指令对应的第一数据，所述第一数据包括输入神经元数据和权值数据，所述第一指令包括排序指令或者稀疏处理指令;

所述运算单元响应所述第一指令，对所述输入神经元数据和所述权值数据执行所述第一指令对应的运算操作，得到运算结果。

22、根据权利要求21所述的方法，其特征在于，所述控制器单元包括:指令缓存单元和指令处理单元，所述指令缓存单元用于对指令进行缓存，所述指令处理单元用于实现译码功能。

23、根据权利要求22所述的方法，其特征在于，所述运算装置还包括:配置解析单元和映射单元，在所述第一指令为所述稀疏处理指令以及所述第一数据还包括预设配置数据时，其中，

所述配置解析单元根据所述预设配置数据设置映射模式;

所述映射单元根据所述映射模式对所述输入神经元和所述权值数据进行映射处理，得到输入神经元-权值对，所述输入神经元-权值对为映射处理后的输入神经元数据与权值数据之间的映射关系;

所述指令缓存单元接收由所述控制器单元发送的目标指令;

所述指令处理单元将所述目标指令译码为运算指令；由所述运算单元对所述输入神经元-权值对执行运算操作，得到运算结果。

24、根据权利要求 21 所述的方法，其特征在于，所述第一数据还包括稀疏参数；所述方法还包括：

稀疏单元依据所述稀疏参数对所述运算结果进行稀疏处理，得到稀疏处理后的运算结果。

25、根据权利要求 24 所述的方法，其特征在于，所述稀疏参数包括稀疏模式；

所述映射单元根据所述映射模式对所述输入神经元和所述权值进行映射处理，具体为：

当处于所述稀疏模式为稀疏模式 1 时，获取所述稀疏模式 1 对应的权值稀疏序列，并依据该权值稀疏序列对所述权值进行映射处理；

当处于所述稀疏模式为稀疏模式 2 时，获取所述稀疏模式 2 对应的神经元稀疏序列，并依据该神经元稀疏序列对所述输入神经元进行映射处理；

当处于所述稀疏模式为稀疏模式 3 时，获取所述稀疏模式 3 对应的权值稀疏序列和神经元稀疏序列，并依据该权值稀疏序列和神经元稀疏序列对所述输入神经元和所述权值数据进行映射处理。

26、根据权利要求 24 或 25 所述的方法，其特征在于，所述稀疏参数还包括稀疏率，所述稀疏单元依据所述稀疏参数对所述运算结果进行稀疏处理，具体为：

对神经元数据的元素绝对值排序，根据稀疏率计算获得需要稀疏的元素个数，根据需要稀疏的元素个数对排序后的神经元数据的元素作稀疏处理，并将稀疏后的稀疏神经元数据和神经元稀疏序列发送至所述控制器单元。

27、根据权利要求 24 或 25 所述的方法，其特征在于，所述稀疏单元依据所述稀疏参数对所述运算结果进行稀疏处理，具体为：

神经元数据为 0 的元素保持 0 不变，神经元数据在预设取值区间内的元素置为 0。

28、根据权利要求 22 所述的方法，其特征在于，如所述输入神经元数据为向量；所述第一指令为向量排序指令，以及所述第一数据为待排序数据向量

以及待排序向量的中间结果；所述控制器单元包括指令处理单元；

所述指令处理单元将所述向量排序指令译码成所述运算单元执行的微指令；

所述运算单元根据所述微指令将所述待排序数据向量或所述中间结果进行排序，得到与所述待排序数据向量等长度的排序后的向量。

29、根据权利要求 28 所述的方法，其特征在于，所述运算单元根据所述微指令将所述待排序数据向量或所述中间结果进行排序具体为：

步骤 A：若排序后得到的为待排序数据向量的中间结果，则将所述待排序数据向量的中间结果写回到存储单元的源地址，并重复执行所述步骤 A，直到所述得到待排序数据向量的最后结果，则跳转到步骤 B；步骤 B：若排序得到的为待排序数据向量的最后结果，将所述待排序数据向量的最后结果根据所述向量排序指令提供的目的操作数地址写回到存储单元的数据 I/O 单元，操作结束。

30、根据权利要求 29 所述的方法，其特征在于，所述运算单元包括 n 个向量归并单元构成，其中，n 为大于等于 2 的整数，所述 n 个向量归并单元用于从所述存储单元中读取不大于 2n 个已经归并的子向量或者有序子向量，并进行归并，将归并后的结果转存入所述存储单元中，直到已经归并的子向量的长度等于所述待排序数据向量长度，形成排序后的向量。

31、根据权利要求 30 所述的方法，其特征在于，所述运算单元执行步骤 A 具体为：

步骤 A1、初始化归并次数 i 为 1；

步骤 A2、由所述 n 个向量归并单元进行计算，在第 i 次归并所述待排序数据向量或所述中间结果时，从所述存储单元中获取所述待排序数据向量或所述中间结果，将所述待排序数据向量或中间结果按顺序分成$\lceil m/2^{i-1} \rceil$份，对向量进行两两归并，除最后一份外，每个向量长度为2^{i-1}；m 为待排序数据向量的长度；

步骤 A3、若归并次数$i < \lceil log_2 m \rceil$，则将归并次数加一，并将处理后的中间结果写回到存储单元源地址，重复执行步骤 A2-A3，直到$i = \lceil log_2 m \rceil$，则跳转到步骤 B;

所述运算单元执行步骤 B，具体为:

若归并次数$i = \lceil log_2 m \rceil$，若只存在分配后的两份待排序数据向量，则经所述 n 个向量归并单元中的第一个向量归并单元归并后，得到的向量为已排序向量，将排序后的结果根据所述向量排序指令提供的目的操作数地址写入到数据输出单元中，操作结束。

32、根据权利要求 31 所述的方法，其特征在于，所述运算单元对向量进行两两归并，具体为:

根据所述向量排序指令提供的源操作数地址按顺序编号 1、2、…、$\lceil m/2^{i-1} \rceil$，将编号为 2*j-1、2*j 的向量分配给第（（j -1）mod n ）+ 1 个向量归并单元进行处理，其中 j>0。

33、根据权利要求 21-32 任一项所述的装置，其特征在于，所述待排序数据向量为预处理阶段测试数据特征矩阵对应的特征值向量和分类结果的概率向量。

34、根据权利要求 21-33 所述的方法，其特征在于，所述第一指令包括下述指令中的一个或任意组合: 向量间与指令 VAV、向量内与指令 VAND、向量间或指令 VOV、向量内或指令 VOR、向量指数指令 VE、向量对数指令 VL、向量大于判定指令 VGT、向量等于判定指令 VEQ、向量非指令 VINV、向量选择合并指令 VMER、向量最大值指令 VMAX、标量扩展指令 STV、标量替换向量指令 STVPN、向量替换标量指令 VPNTS、向量检索指令 VR、向量点积指令 VP、随机向量指令 RV、循环移位指令 VCS、向量加载指令 VLOAD、向量存储指令 VS、向量搬运指令 VMOVE、矩阵检索指令 MR、矩阵加载指令 ML、矩阵存储指令 MS、矩阵搬运指令 MMOVE。

35、根据权利要求 21-34 任一项所述的方法，其特征在于，所述装置用于稀疏神经网络运算或者稠密神经网络运算。

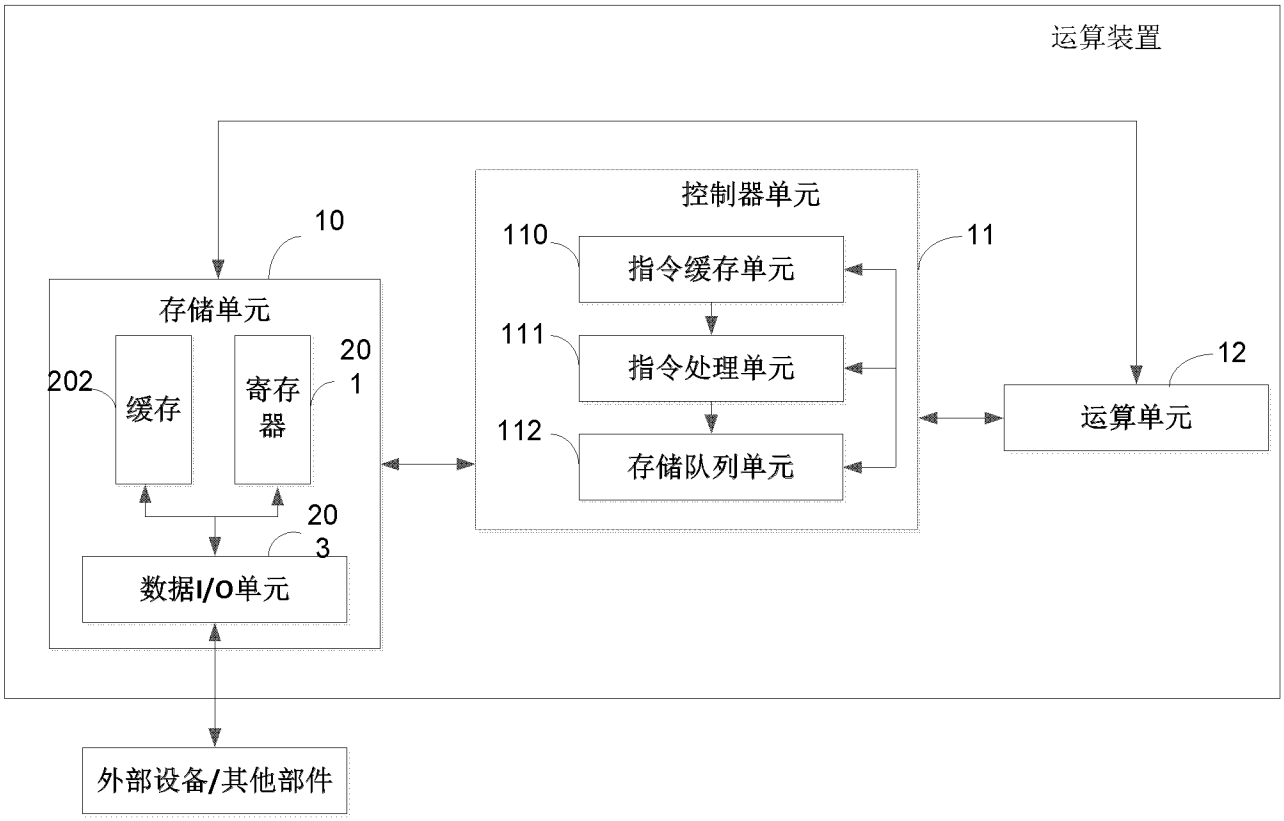

运算装置
控制器单元
10
110
指令缓存单元
11
存储单元
111
指令处理单元
12
202
缓存
寄存器
201
112
存储队列单元
运算单元
203
数据I/O单元
外部设备/其他部件

图 1A

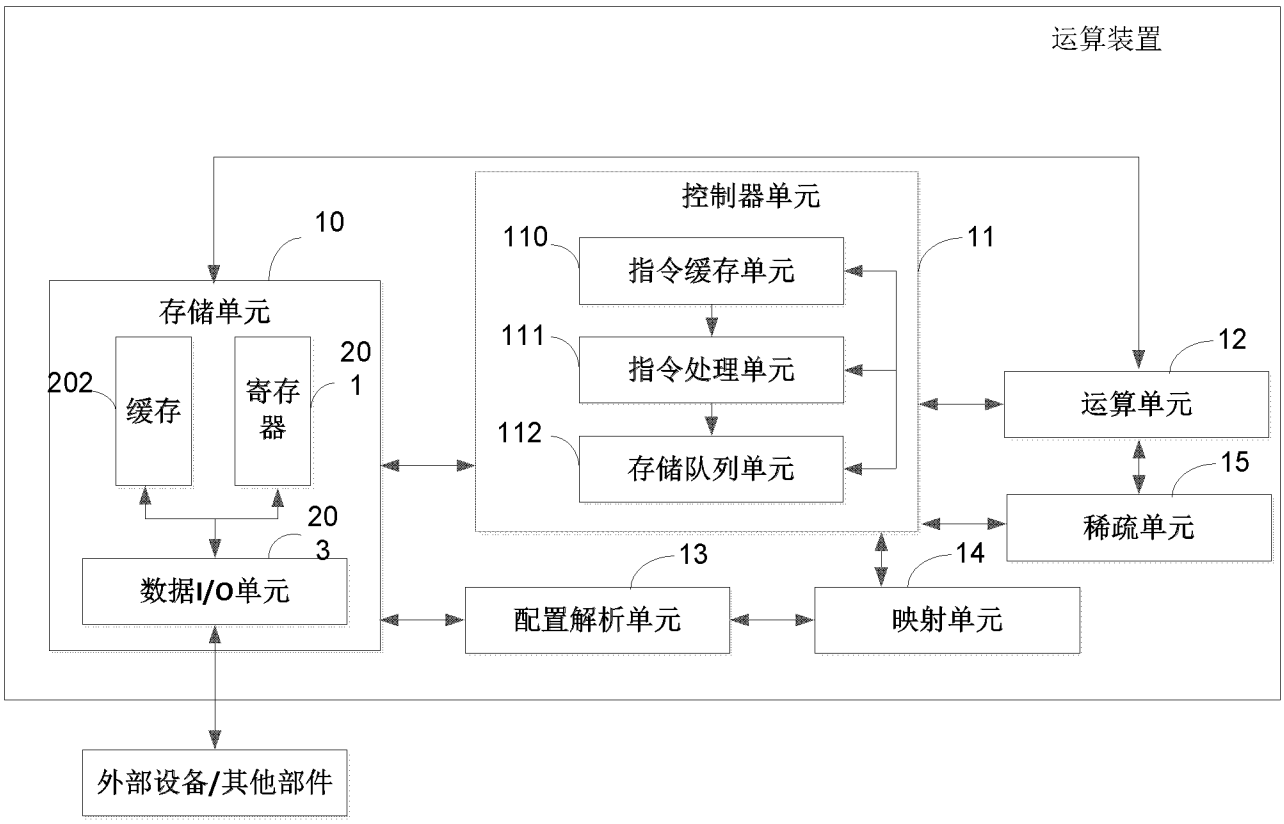

运算装置
控制器单元
10
110
指令缓存单元
11
存储单元
111
指令处理单元
12
202
缓存
寄存器
20
1
112
存储队列单元
运算单元
15
稀疏单元
20
3
13
14
数据I/O单元
配置解析单元
映射单元
外部设备/其他部件

图 1B

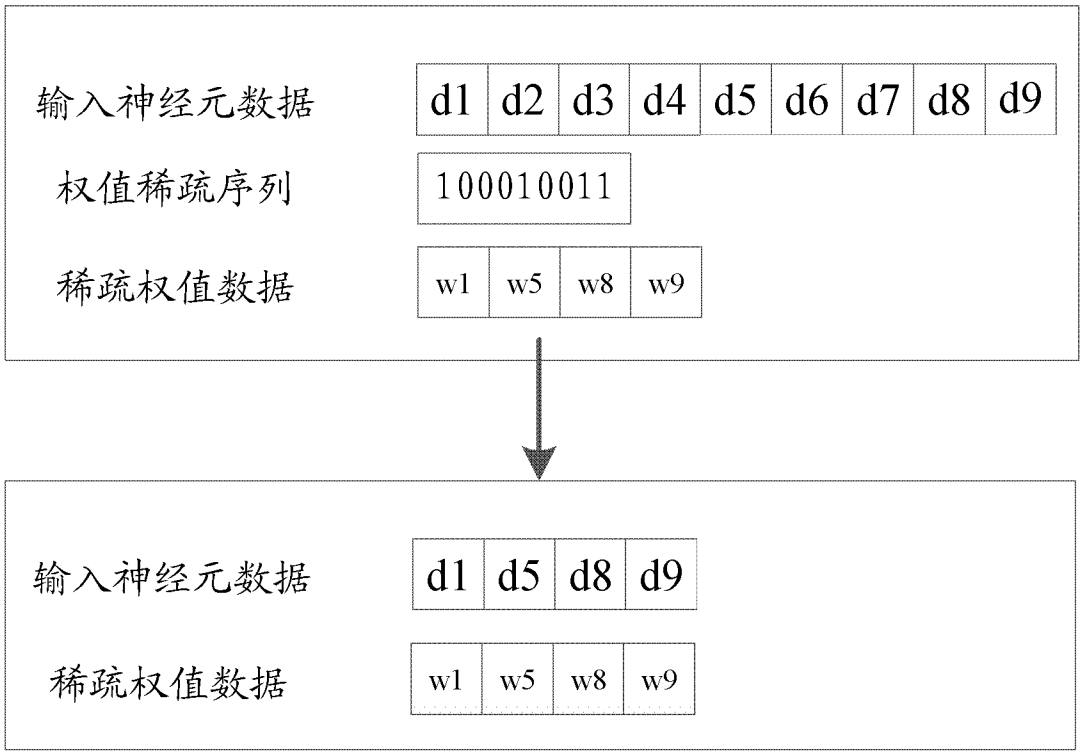

输入神经元数据
d1 d2 d3 d4 d5 d6 d7 d8 d9
权值稀疏序列
100010011
稀疏权值数据
w1 w5 w8 w9
输入神经元数据
d1 d5 d8 d9
稀疏权值数据
w1 w5 w8 w9

图 1C

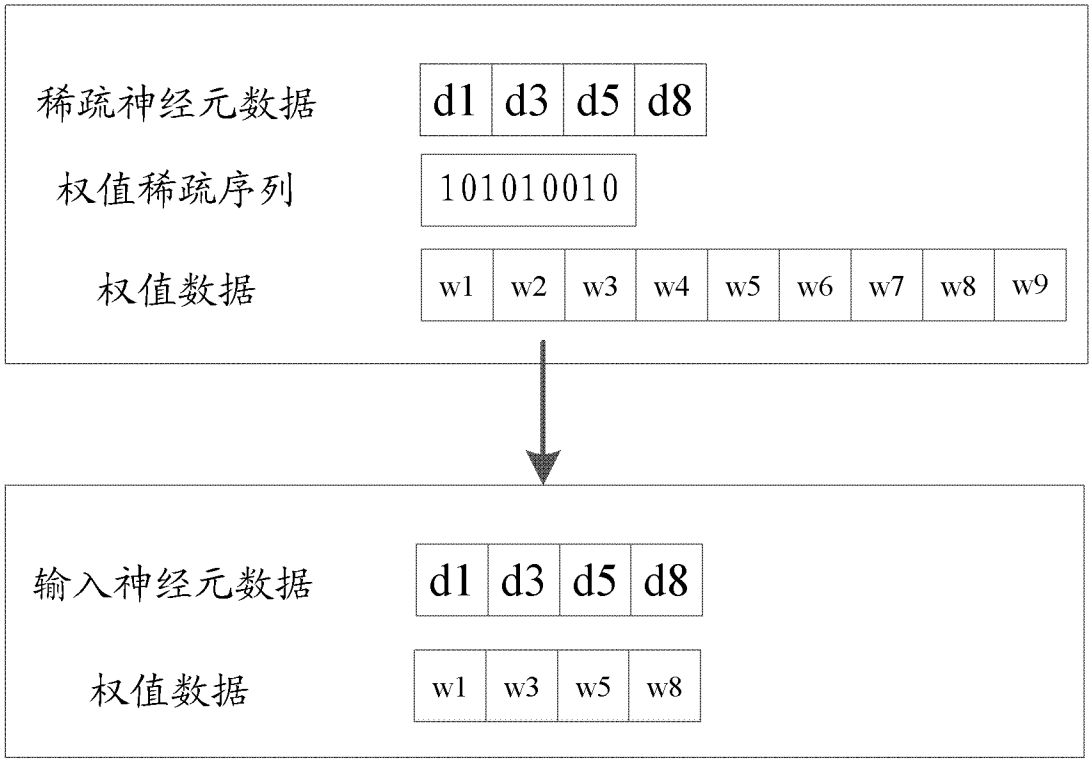

稀疏神经元数据
d1 d3 d5 d8
权值稀疏序列
101010010
权值数据
w1 w2 w3 w4 w5 w6 w7 w8 w9
输入神经元数据
d1 d3 d5 d8
权值数据
w1 w3 w5 w8

图 1D

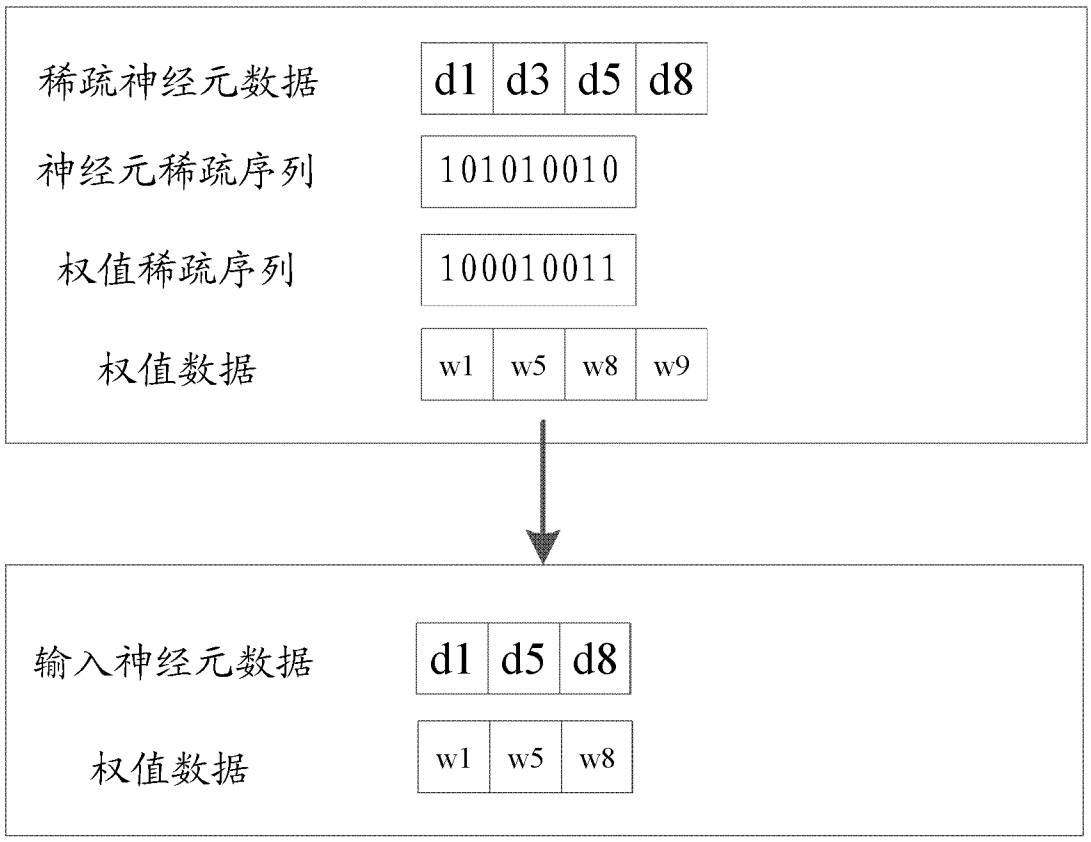

稀疏神经元数据
d1 d3 d5 d8
神经元稀疏序列
101010010
权值稀疏序列
100010011
权值数据
w1 w5 w8 w9
输入神经元数据
d1 d5 d8
权值数据
w1 w5 w8

图 1E

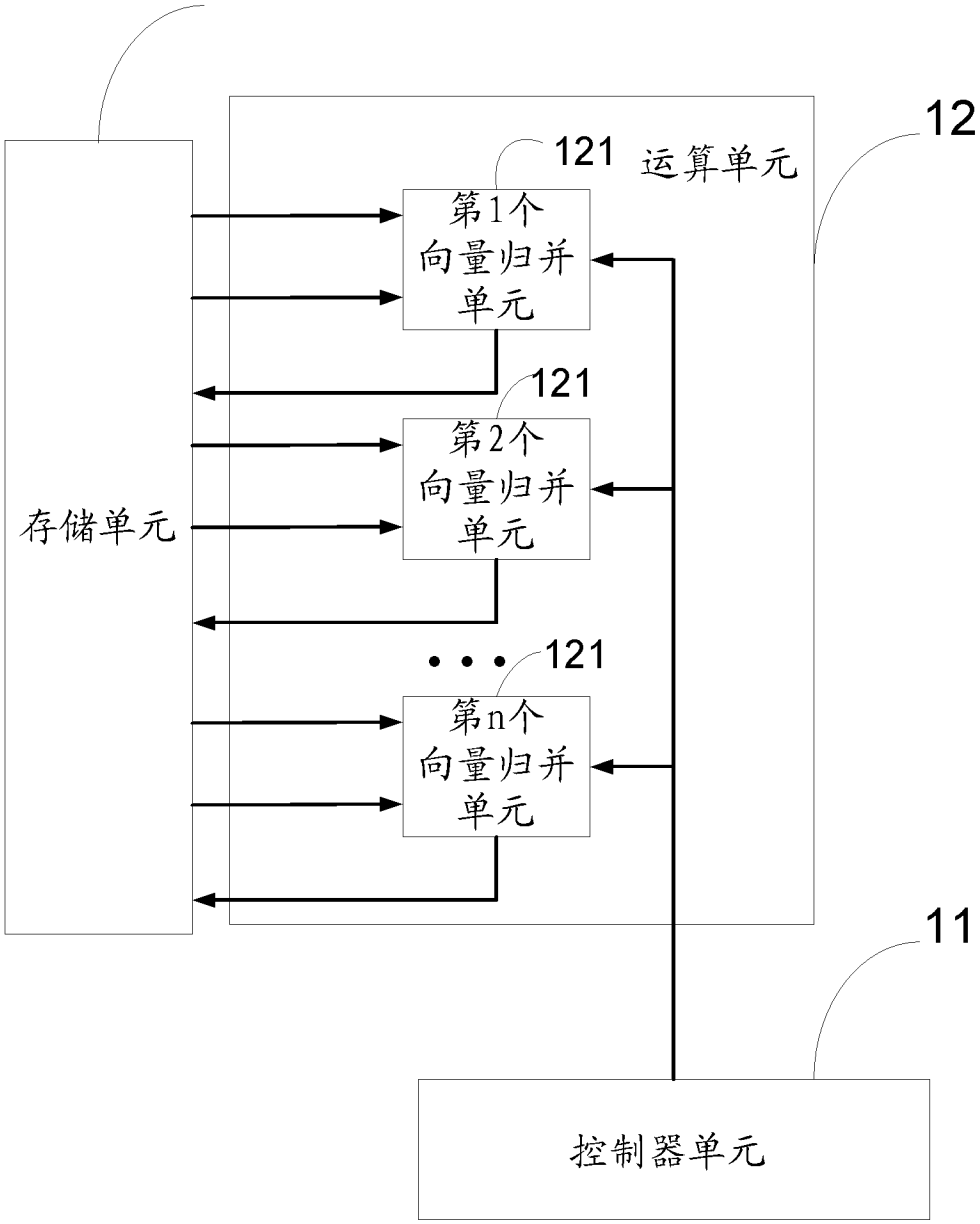

12
121
运算单元
第1个
向量归并
单元
121
第2个
向量归并
单元
存储单元
121
第n个
向量归并
单元
11
控制器单元

图 1F

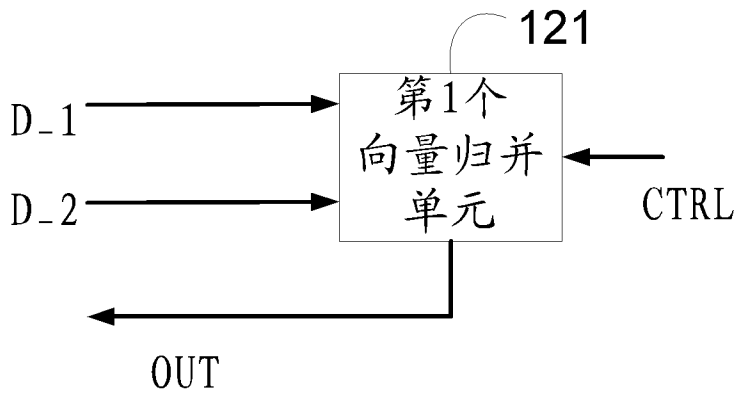
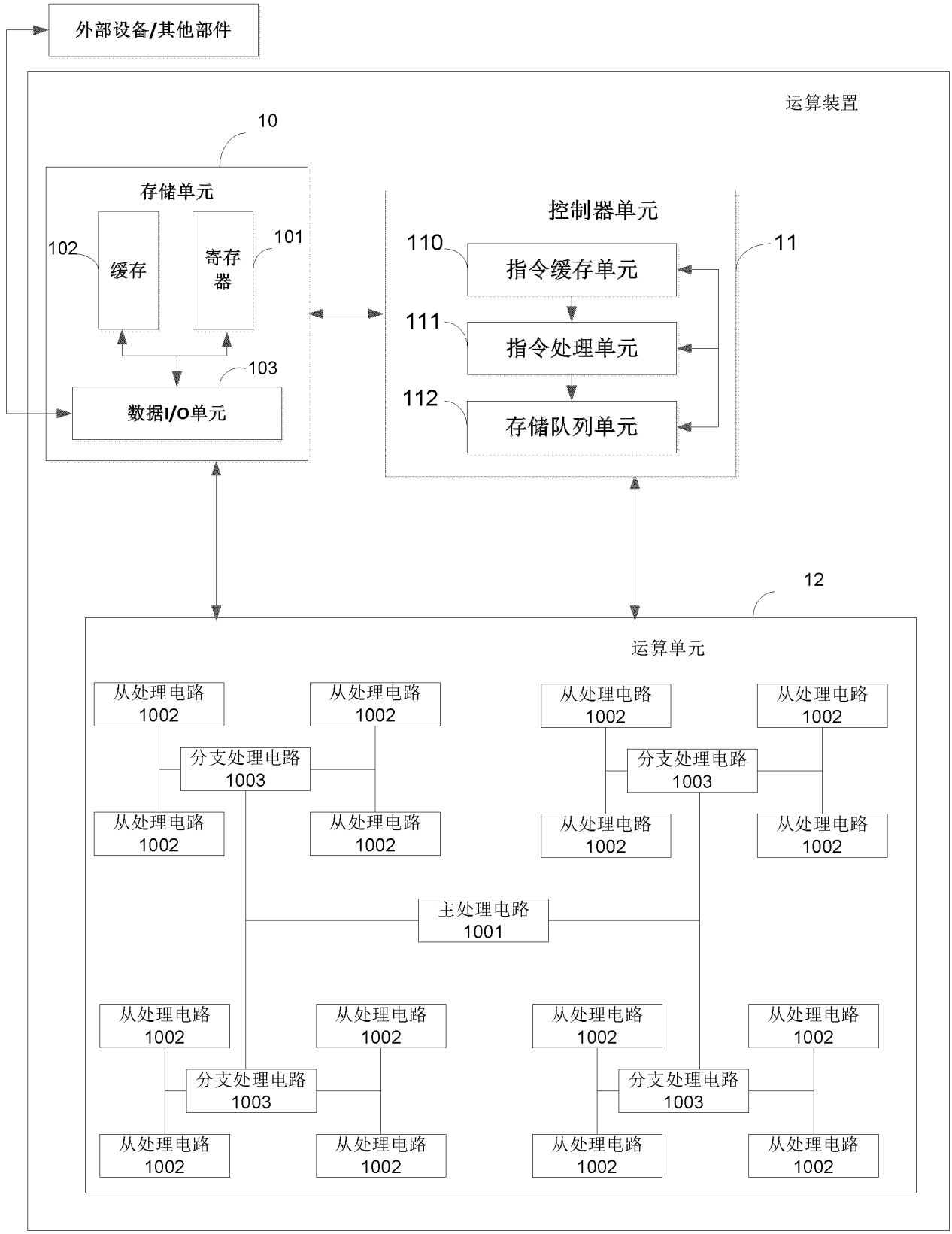

121
第1个
向量归并
单元
D_1
D_2
CTRL
OUT

图 1G


外部设备/其他部件
运算装置
10
存储单元
102
缓存
寄存器
101
103
数据I/O单元
控制器单元
110
指令缓存单元
11
111
指令处理单元
112
存储队列单元
12
运算单元
从处理电路
1002
分支处理电路
1003
主处理电路
1001

1H

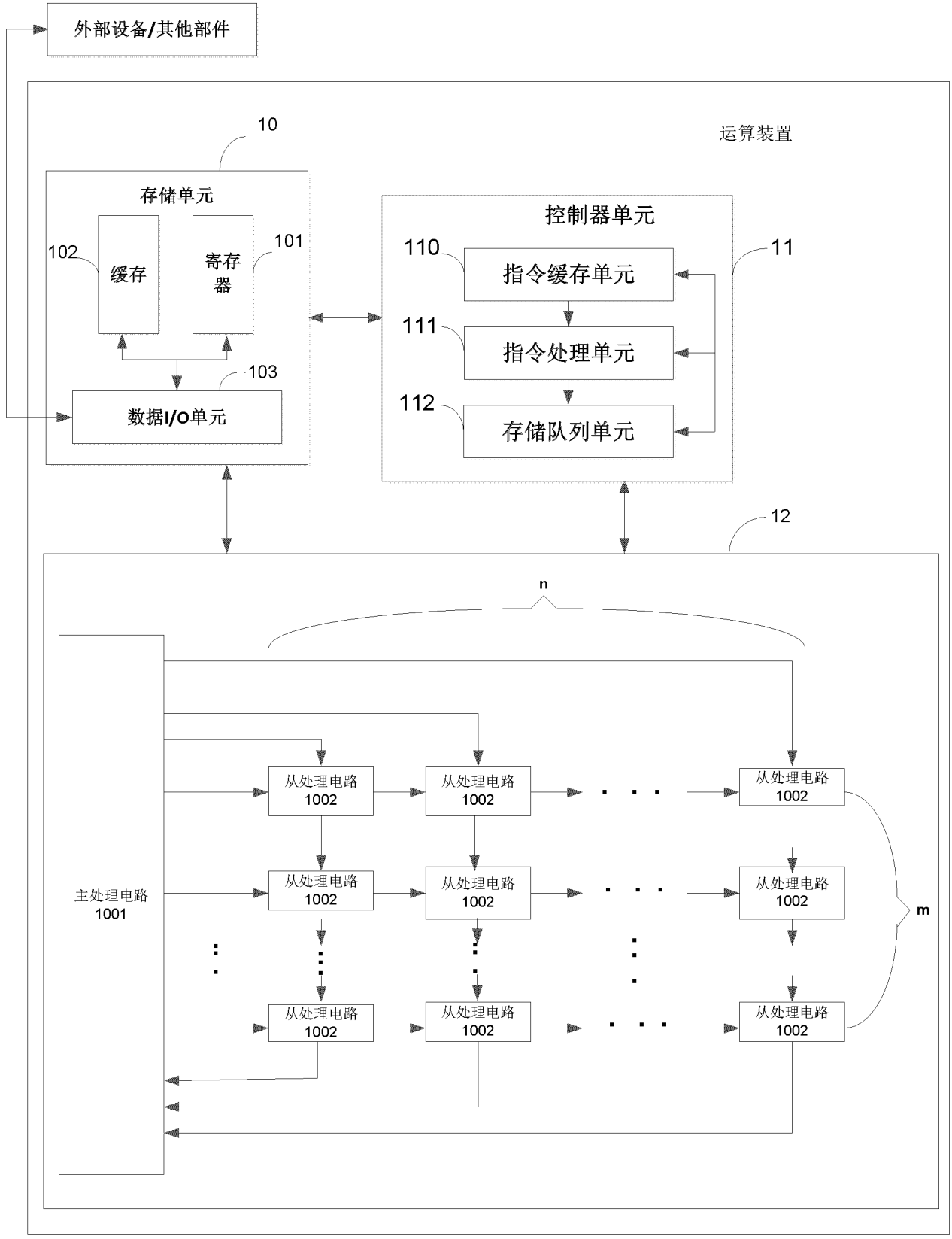

外部设备/其他部件
运算装置
10
存储单元
102
缓存
寄存器
101
103
数据I/O单元
控制器单元
110
指令缓存单元
11
111
指令处理单元
112
存储队列单元
12
n
主处理电路
1001
从处理电路
1002
m

图 1I


主处理电路1001
转换处理电路1010
激活处理电路1011
加法处理电路1012

图 1J

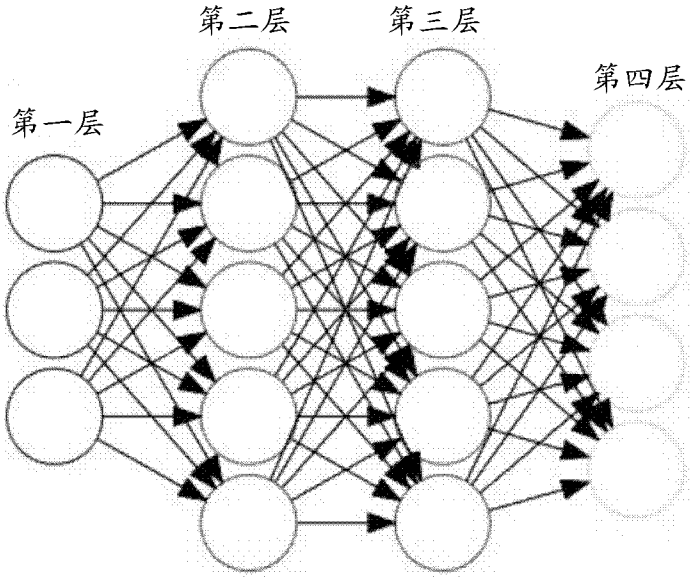

第一层
第二层
第三层
第四层

图 1K

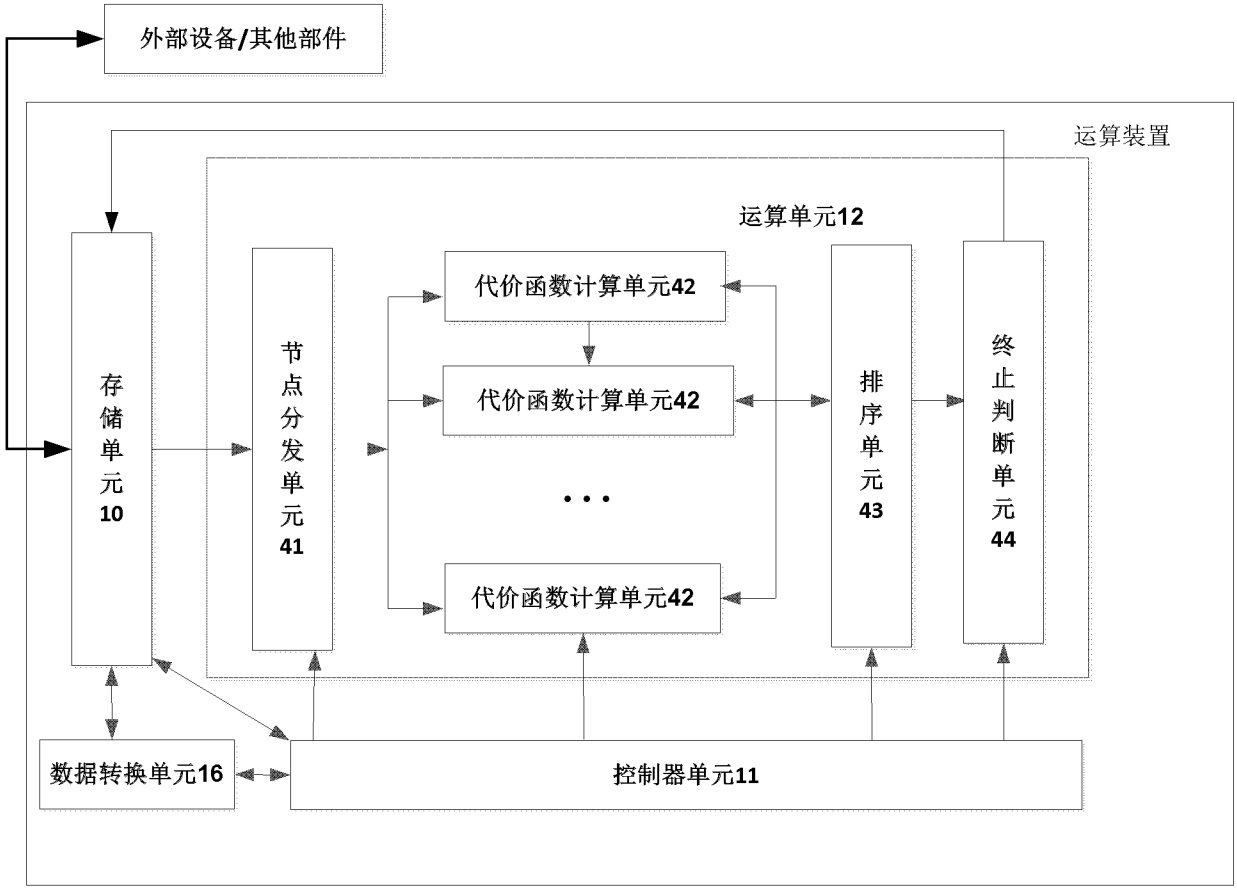

外部设备/其他部件
运算装置
运算单元12
存储单元10
节点分发单元41
代价函数计算单元42
代价函数计算单元42
代价函数计算单元42
排序单元43
终止判断单元44
数据转换单元16
控制器单元11

图 1L

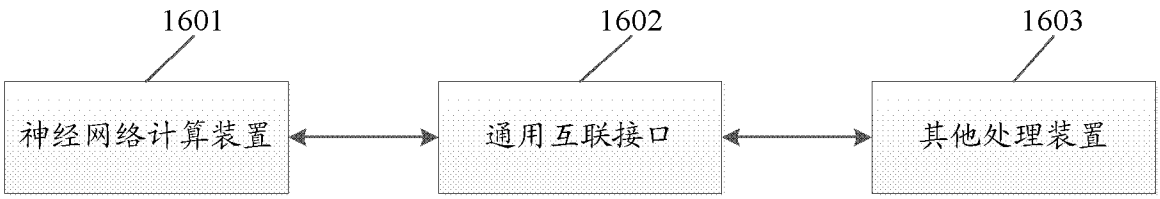

1601
1602
1603
神经网络计算装置
通用互联接口
其他处理装置

图 2A

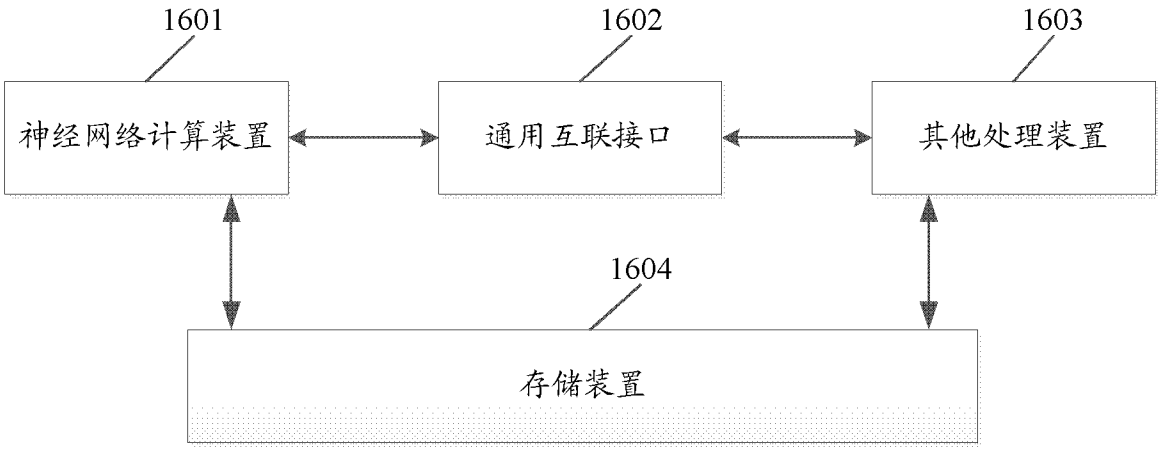


图 2B

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2018/083379

A. CLASSIFICATION OF SUBJECT MATTER

G06N 3/08 (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 3/-; G6N 5/-; G06N 7/-

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)

USTXT; CNABS; SIPOABS; DWPI; CNKI: spars+, nearal network, neuron, deep learning, weight, sequence, sort, rank, order, memory, stor+, vector, instruction, command, map+, 稀疏, 神经网络, 神经元, 深度学习, 权重, 权值, 序列, 排序, 存储, 向量, 指令, 命令, 映射

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2017061328 A1 (QUALCOMM INC.) 02 March 2017 (02.03.2017), description, paragraphs [0029], [0030], [0061]-[0063], [0068] and [0069]	1, 2, 4, 6, 7, 13-22, 24, 26, 27, 33-35
X	WO 2016154440 A1 (HRL LAB LLC. et al.) 29 September 2016 (29.09.2016), description, paragraphs [0015]-[0024]	1, 2, 4, 6, 7, 13-22, 24, 26, 27, 33-35
A	US 2016379111 A1 (MICROSOFT TECHNOLOGY LICENSING LLC.) 29 December 2016 (29.12.2016), entire document	1-35
A	WO 2016160237 A1 (QUALCOMM INC.) 06 October 2016 (06.10.2016), entire document	1-35

☐ 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
16 July 2018	23 July 2018
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 WU, Guangping Telephone No. (86-10) 62411829

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2018/083379

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
US 2017061328 A1	02 March 2017	WO 2017039946 A1	09 March 2017
		CA 2993011 A1	09 March 2017
		CN 107924486 A	17 April 2018
		KR 20180048930 A	10 May 2018
		EP 3274930 A1	31 January 2018
		US 2017316311 A1	02 November 2017
US 2016379111 A1	29 December 2016	CN 107735803 A	23 February 2018
		WO 2016210014 A1	29 December 2016
		EP 3314543 A1	02 May 2018
WO 2016160237 A1	06 October 2016	US 2016283864 A1	29 September 2016
		EP 3274927 A1	31 January 2018
		JP 2018514852 A	07 June 2018
		CN 107430703 A	01 December 2017

国际检索报告

国际申请号 PCT/CN2018/083379

A. 主题的分类

G06N 3/08(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 3/-, G6N 5/-, G06N 7/-

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））

USTXT;CNABS;SIPOABS;DWPI;CNKI:spars+, nearal network, neuron, deep learning, weight, sequence, sort, rank, order,memory, stor+, vector, instruction, command, map+, 稀疏，神经网络，神经元，深度学习，权重，权值，序列，排序，存储，向量，指令，命令，映射

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	US 2017061328 A1 (QUALCOMM INC) 2017年 3月 2日 (2017-03-02) 说明书第[0029]-[0030]、[0061]-[0063]、[0068]-[0069]段	1-2，4，6-7，13-22，24，26-27，33-35
X	WO 2016154440 A1 (HRL LAB LLC等) 2016年 9月 29日 (2016-09-29) 说明书第[0015]-[0024]段	1-2，4，6-7，13-22，24，26-27，33-35
A	US 2016379111 A1 (MICROSOFT TECHNOLOGY LICENSING LLC) 2016年 12月 29日 (2016-12-29) 全文	1-35
A	WO 2016160237 A1 (QUALCOMM INC) 2016年 10月 6日 (2016-10-06) 全文	1-35

☐ 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
2018年 7月 16日	2018年 7月 23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 吴广平 电话号码 62411829

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2018/083379

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）
US	2017061328	A1	2017年 3月 2日	WO	2017039946	A1	2017年 3月 9日
				CA	2993011	A1	2017年 3月 9日
				CN	107924486	A	2018年 4月 17日
				KR	20180048930	A	2018年 5月 10日
WO	2016154440	A1	2016年 9月 29日	CN	107251059	A	2017年 10月 13日
				EP	3274930	A1	2018年 1月 31日
				US	2017316311	A1	2017年 11月 2日
US	2016379111	A1	2016年 12月 29日	CN	107735803	A	2018年 2月 23日
				WO	2016210014	A1	2016年 12月 29日
				EP	3314543	A1	2018年 5月 2日
WO	2016160237	A1	2016年 10月 6日	US	2016283864	A1	2016年 9月 29日
				EP	3274927	A1	2018年 1月 31日
				JP	2018514852	A	2018年 6月 7日
				CN	107430703	A	2017年 12月 1日