(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 109902810 B
(45)授权公告日 2020.05.22

(21)申请号 201711315639.9

(22)申请日 2017.12.11

(65)同一申请的已公布的文献号
申请公布号 CN 109902810 A

(43)申请公布日 2019.06.18

(73)专利权人 中科寒武纪科技股份有限公司
地址 100000 北京市海淀区科学院南路6号科研综合楼644室

(72)发明人 不公告发明人

(74)专利代理机构 广州三环专利商标代理有限公司 44202
代理人 郝传鑫 熊永强

(51)Int.Cl.
G06N 3/06(2006.01)

(56)对比文件
CN 105512723 A,2016.04.20,
CN 106779068 A,2017.05.31,
US 8983885 B1,2015.03.17,

审查员 姚希

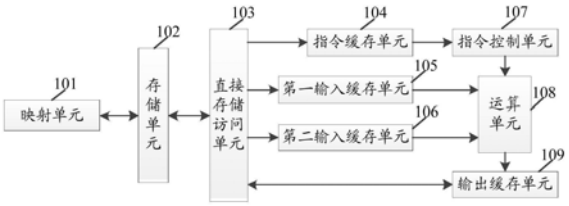
权利要求书7页 说明书31页 附图10页

(54)发明名称

神经网络运算设备和方法

(57)摘要

本发明公开了一种神经网络处理模块，其映射单元接收到输入神经元和权值后，对该输入神经元和/或权值进行处理，以得到处理后的输入神经元和处理后的权值；神经网络处理模块的运算单元对处理后的输入神经元和处理后的权值进行人工神经网络运算。采用本发明实施例可减少装置的额外开销，并减小访问量，提高了神经网络运算效率。



CN 109902810 B

1.一种神经网络运算模块,其特征在于,包括:

映射单元,用于接收输入数据之后,对所述输入数据进行处理,以得到处理后的输入数据,所述输入数据包括至少一个输入神经元和至少一个权值,所述处理后的输入数据包括处理后的输入神经元和处理后的权值;

其中,所述映射单元包括:

第一稀疏处理单元,用于对第二输入数据进行处理,以得到第三输出数据和第二输出数据,并将所述第三输出数据传输至第一数据处理单元;

第一数据处理单元,用于根据所述第三输出数据对第一输入数据进行处理,以得到第一输出数据;

其中,当所述第一输入数据包括至少一个输入神经元,所述第二输入数据包括至少一个权值时,所述第一输出数据为处理后的输入神经元,所述第二输出数据为处理后的权值,所述第三输出数据为权值的连接关系数据;当所述第一输入数据包括至少一个权值,所述第二输入数据包括至少一个输入神经元时,所述第一输出数据为处理后的权值,所述第二输出数据为处理后的输入神经元,所述第三输出数据为输入神经元的连接关系数据;

或者,

所述映射单元包括:

第二稀疏处理单元,用于接收到第三输入数据后,根据所述第三输入数据得到第一连接关系数据,并将该第一连接关系数据传输至连接关系处理单元;

第三稀疏处理单元,用于接收到第四输入数据后,根据所述第四输入数据得到第二连接关系数据,并将该第二连接关系数据传输至所述连接关系处理单元;

所述连接关系处理单元,用于根据所述第一连接关系数据和所述第二连接关系数据,以得到第三连接关系数据,并将该第三连接关系数据传输至第二数据处理单元;

所述第二数据处理单元,用于在接收到所述第三输入数据,所述第四输入数据和所述第三连接关系数据后,根据所述第三连接关系数据对所述第三输入数据和所述第四输入数据进行处理,以得到第四输出数据和第五输出数据;

其中,当所述第三输入数据包括至少一个输入神经元,第四输入数据包括至少一个权值时,所述第一连接关系数据为输入神经元的连接关系数据,所述第二连接关系数据为权值的连接关系数据,所述第四输出数据为处理后的输入神经元,所述第五输出数据为处理后的权值;当所述第三输入数据包括至少一个权值,所述第四输入数据包括至少一个输入神经元时,所述第一连接关系数据为权值的连接关系数据,所述第二连接关系数据为输入神经元的连接关系数据,所述第四输出数据为处理后的权值,所述第五输出数据为处理后的输入神经元;

存储单元,用于存储所述处理后的输入神经元、处理后的权值、神经网络指令和运算结果;

指令控制单元,用于从指令缓存单元中获取所述神经网络指令,并将所述神经网络指令译码成运算单元执行的微指令;

所述运算单元,用于从所述第一输入缓存单元和所述第二输入缓存单元中获取所述处理后的输入神经元和所述处理后的权值后,根据所述微指令对所述处理后的输入神经元和所述处理后的权值进行人工神经网络运算,以得到所述运算结果;

输出缓存单元，用于缓存所述运算结果；

当所述第一连接关系数据和所述第二连接关系数据均以步长索引的形式表示，且表示所述第一连接关系数据和所述第二连接关系数据的字符串是按照物理地址由低到高的顺序存储时，所述连接关系处理单元具体用于：

将所述第一连接关系数据的字符串中的每一个元素与存储物理地址低于该元素存储的物理地址的元素进行累加，得到的新的元素组成第四连接关系数据；同理，对所述第二连接关系数据的字符串进行同样的处理，得到第五连接关系数据；

从所述第四连接关系数据的字符串和所述第五连接关系数据的字符串中，选取相同的元素，按照元素值从小到大的顺序排序，组成新的字符串；

将所述新的字符串中每一个元素与其相邻的且值小于该元素值的元素进行相减，得到的元素组成所述第三连接关系数据；

当所述第一连接关系数据和所述第二连接关系数据均以直接索引的形式表示时，所述连接关系处理单元具体用于：

对所述第一连接关系数据和所述第二连接关系数据进行与操作，以得到第三连接关系数据；

当所述第一连接关系数据与所述第二连接关系数据中任意一个以步长索引的形式表示，另一个以直接索引的形式表示时，所述连接关系处理单元具体用于：

若所述第一连接关系数据是以步长索引的形式表示，将所述第一连接关系数据转换成以直接索引的形式表示的连接关系数据；

若所述第二连接关系数据是以步长索引的形式表示，将所述第二连接关系数据转换成以直接索引的形式表示的连接关系数据；

对所述第一连接关系数据和所述第二连接关系数据进行与操作，以得到第三连接关系数据；

当所述第一连接关系数据与所述第二连接关系数据中任意一个以步长索引的形式表示，另一个以直接索引的形式表示，且表示所述第一连接关系数据和所述第二连接关系数据的字符串是按照物理地址由低到高的顺序存储时，所述连接关系处理单元还具体用于：

若所述第一连接关系数据是以步长索引的形式表示，将所述第二连接关系数据转换成以步长索引的形式表示的连接关系数据；

若所述第二连接关系数据是以步长索引的形式表示，将所述第一连接关系数据转换成以步长索引的形式表示的连接关系数据；

将所述第一连接关系数据的字符串中的每一个元素与存储物理地址低于该元素存储的物理地址的元素进行累加，得到的新的元素组成第四连接关系数据；同理，对所述第二连接关系数据的字符串进行同样的处理，得到第五连接关系数据；

从所述第四连接关系数据的字符串和所述第五连接关系数据的字符串中，选取相同的元素，按照元素值从小到大的顺序排序，组成新的字符串；

将所述新的字符串中每一个元素与其相邻的且值小于该元素值的元素进行相减，得到的元素组成所述第三连接关系数据。

2.根据权利要求1所述的神经网络运算模块，其特征在于，所述神经网络运算模块还包括：

直接存储访问单元，用于在所述存储单元与所述指令缓存单元、所述第一输入缓存单元、所述第二输入缓存单元和所述输出缓存单元之间进行数据的读写；

所述指令缓存单元，用于缓存所述直接存储访问单元读取所述神经网络指令；

所述第一输入缓存单元，用于缓存所述直接存储访问单元读取的第一缓存数据，所述第一缓存数据为所述处理后的输入神经元或所述处理后的权值；

所述第二输入缓存单元，用于缓存所述直接存储访问单元读取的第二缓存数据，所述第二缓存数据为所述处理后的权值或所述处理后的输入神经元，且所述第一缓存数据与所述第二缓存数据不一致。

3.根据权利要求1或2所述的神经网络运算模块，其特征在于，所述输入神经元的连接关系数据和所述权值的连接关系数据均以直接索引或者步长索引的形式表示；

当所述输入神经元的连接关系数据以直接索引的形式表示时，该连接关系数据为由0和1组成的字符串，0表示所述输入神经元的绝对值小于或者等于第一阈值，1表示所述输入神经元的绝对值大于所述第一阈值；

当所述输入神经元的连接关系数据以步长索引形式表示时，该连接关系数据为绝对值大于所述第一阈值的输入神经元与上一个绝对值大于所述第一阈值的输入神经元之间的距离值组成的字符串；

当所述权值的连接关系数据以直接索引的形式表示时，该连接关系数据为由0和1组成的字符串，0表示所述权值的绝对值小于或者等于第二阈值，即该权值对应的输入神经元与输出神经元之间没有连接，1表示所述权值的绝对值大于所述第二阈值，即该权值对应的输入神经元与输出神经元之间有连接；以直接索引形式表示权值的连接关系数据有两种表示顺序：以每个输出神经元与所有输入神经元的连接状态组成一个0和1的字符串来表示所述权值的连接关系数据；或者每个输入神经元与所有输出神经元的连接状态组成一个0和1的字符串来表示所述权值的连接关系数据；

当所述权值的连接关系数据以步长索引的形式表示时，该连接关系数据为与输出神经元有连接的输入神经元的与上一个与该输出神经元有连接的输入神经元之间的距离值组成的字符串。

4.根据权利要求1或2所述神经网络运算模块，其特征在于，所述映射单元对所述输入数据进行处理之前，所述映射单元还用于：

对所述至少一个输入神经元进行分组，以得到M组输入神经元，所述M为大于或者等于1的整数；

判断所述M组输入神经元的每一组输入神经元是否满足第一预设条件，所述第一预设条件包括一组输入神经元中绝对值小于或者等于第三阈值的输入神经元的个数小于或者等于第四阈值；

当所述M组输入神经元任意一组输入神经元不满足所述第一预设条件时，将该组输入神经元删除；

对所述至少一个权值进行分组，以得到N组权值，所述N为大于或者等于1的整数；

判断所述N组权值的每一组权值是否满足第二预设条件，所述第二预设条件包括一组权值中绝对值小于或者等于第五阈值的权值的个数小于或者等于第六阈值；

当所述N组权值任意一组权值不满足所述第二预设条件时，将该组权值删除。

5.根据权利要求1-2任一项所述的神经网络运算模块，其特征在于，所述神经网络运算模块用于稀疏神经网络运算或者稠密神经网络运算。

6.一种神经网络运算装置，其特征在于，所述神经网络运算装置包括一个或多个如权利要求1-5任一项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；

当所述神经网络运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，多个所述神经网络运算模块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；多个所述神经网络运算模块共享同一控制系统或拥有各自的控制系统；多个所述神经网络运算模块共享内存或者拥有各自的内存；多个所述神经网络运算模块的互联方式是任意互联拓扑。

7.一种组合处理装置，其特征在于，所述组合处理装置包括如权利要求6所述的神经网络运算装置，通用互联接口和其他处理装置；

所述神经网络运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。

8.一种神经网络芯片，其特征在于，所述神经网络芯片包括如权利要求6所述的神经网络运算装置或如权利要求7所述的组合处理装置。

9.一种板卡，其特征在于，所述板卡包括如权利要求8所述的神经网络芯片。

10.一种电子装置，其特征在于，所述电子装置包括如权利要求8所述的神经网络芯片或者如权利要求9所述的板卡。

11.一种神经网络运算方法，其特征在于，包括：

对输入数据进行处理，以得到处理后的输入数据；

其中，所述输入数据包括第一输入数据和第二输入数据，所述处理后的输入数据包括处理后的第一输入数据和处理后的第二输入数据，所述对输入数据进行处理，以得到处理后的输入数据，包括：

对所述第二输入数据进行处理，以得到第一连接关系数据和处理后的第二输出数据；根据所述第一连接关系数据对所述第一输入数据进行处理，以得到处理后的第二输入数据，其中，当所述第一输入数据为输入神经元，所述第二输入数据为权值时，所述第一连接关系数据为所述权值的连接关系数据；当所述第一输入数据为权值，所述第二输入数据为输入神经元时，所述第一连接关系数据为输入神经元的连接关系数据；

或者，

所述输入数据包括输入神经元和权值，所述处理后的输入数据包括处理后的输入神经元和处理后的权值，所述对输入数据进行处理，以得到处理后的输入数据，包括：

根据所述输入神经元和所述权值获取所述输入神经元的连接关系数据和所述权值的连接关系数据；对所述输入神经元的连接关系数据和所述权值的连接关系数据进行处理，以得到第二连接关系数据，根据所述第二连接关系数据对所述输入神经元和所述权值进行处理，以得到所述处理后的输入神经元和所述处理后的权值；

或者

所述对输入数据进行处理，以得到处理后的输入数据，包括：

当所述输入数据包括输入神经元和所述输入神经元的连接关系数据时，根据所述输入

神经元的连接关系数据对所述输入神经元进行处理，以得到处理后的输入神经元；

当所述输入数据包括权值和所述权值的连接关系数据时，根据所述权值的连接关系数据对所述权值进行处理，以得到处理后的权值；

获取神经运算指令，将所述神经运算指令译码成微指令；

根据所述微指令对所述处理后的输入数据进行人工神经网络运算，以得到运算结果；

所述输入神经元的连接关系数据和所述权值的连接关系数据以直接索引的形式表示，所述对所述输入神经元的连接关系数据和所述权值的连接关系数据进行处理，以得到第二连接关系数据，包括：

对所述输入神经元的连接关系数据和所述权值的连接关系数据进行与操作，以得到所述第二连接关系数据；

其中，

所述对所述输入神经元的连接关系数据和所述权值的连接关系数据进行处理，以得到第二连接关系数据，包括：

当所述输入神经元的连接关系数据以直接索引的形式表示，所述权值的连接关系数据以步长索引的形式表示时，将所述权值的连接关系数据转换成以直接索引的形式表示的连接关系数据；

当所述权值的连接关系数据以直接索引的形式表示，所述输入神经元的连接关系数据以步长索引的形式表示时，将所述输入神经元的连接关系数据转换成以直接索引的形式表示的连接关系数据；

对所述输入神经元的连接关系数据和所述权值的连接关系数据进行与操作，以得到所述第二连接关系数据；

或者；

当所述输入神经元的连接关系数据和所述权值的连接关系数均以步长的形式表，且表示所述权值的连接关系数据和所述输入神经元的连接关系数据的字符串是按照物理地址由低到高的顺序存储时，所述对所述输入神经元的连接关系数据和所述权值的连接关系数据进行处理，以得到第二连接关系数据，包括：

将所述输入神经元的连接关系数据的字符串中的每一个元素与存储物理地址低于该元素存储的物理地址的元素进行累加，得到的新的元素组成第三连接关系数据；同理，对所述权值的连接关系数据的字符串进行同样的处理，得到第四连接关系数据；

从所述第三连接关系数据的字符串和所述第四连接关系数据的字符串中，选取相同的元素，按照元素值从小到大的顺序排序，组成新的字符串；

将所述新的字符串中每一个元素与其相邻且值小于该元素值的元素进行相减，得到的元素组成所述第二连接关系数据；

或者；

表示所述权值的连接关系数据和所述输入神经元的连接关系数据的字符串是按照物理地址由低到高的顺序存储，所述对所述输入神经元的连接关系数据和所述权值的连接关系数据进行处理，以得到第二连接关系数据，包括：

当所述输入神经元的关系数据是以步长索引的形式表示，所述权值的连接关系数据是以直接索引的形式表示时，将所述权值的连接关系数据转换成以步长索引的形式表示的连

接关系数据；

当所述权值的关系数据是以步长索引的形式表示，所述输入神经元的连接关系数据是以直接索引的形式表示时，将所述输入神经元的连接关系数据转换成以步长索引的形式表示的连接关系数据；

将所述输入神经元的连接关系数据的字符串中的每一个元素与存储物理地址低于该元素存储的物理地址的元素进行累加，得到的新的元素组成第三连接关系数据；同理，对所述权值的连接关系数据的字符串进行同样的处理，得到第四连接关系数据；

从所述第三连接关系数据的字符串和所述第四连接关系数据的字符串中，选取相同的元素，按照元素值从小到大的顺序排序，组成新的字符串；

将所述新的字符串中每一个元素与其相邻且值小于该元素值的元素进行相减，得到的元素组成所述第二连接关系数据。

12.根据权利要求11所述的方法，其特征在于，所述输入数据包括至少一个输入神经元和/或至少一个权值，所述对输入数据进行处理之前，所述方法还包括：

对所述至少一个输入神经元进行分组，以得到M组输入神经元，所述M为大于或者等于1的整数；

判断所述M组输入神经元的每一组输入神经元是否满足第一预设条件，所述第一预设条件包括一组输入神经元中绝对值小于或者等于第三阈值的输入神经元的个数小于或者等于第四阈值；

当所述M组输入神经元任意一组输入神经元不满足所述第一预设条件时，将该组输入神经元删除；

对所述至少一个权值进行分组，以得到N组权值，所述N为大于或者等于1的整数；

判断所述N组权值的每一组权值是否满足第二预设条件，所述第二预设条件包括一组权值中绝对值小于或者等于第五阈值的权值的个数小于或者等于第六阈值；

当所述N组权值任意一组权值不满足所述第二预设条件时，将该组权值删除。

13.根据权利要求11或12所述的方法，其特征在于，所述输入神经元的连接关系数据和所述权值的连接关系数据以直接索引或者步长索引的形式表示；

当所述输入神经元的连接关系数据以直接索引的形式表示时，该连接关系数据为由0和1组成的字符串，0表示所述输入神经元的绝对值小于或者等于第一阈值，1表示所述输入神经元的绝对值大于所述第一阈值；

当所述输入神经元的连接关系数据以步长索引形式表示时，该连接关系数据为绝对值大于所述第一阈值的输入神经元与上一个绝对值大于所述第一阈值的输入神经元之间的距离值组成的字符串；

当所述权值的连接关系数据以直接索引的形式表示时，该连接关系数据为由0和1组成的字符串，0表示所述权值的绝对值小于或者等于第二阈值，即所述权值对应的输入神经元与输出神经元之间没有连接，1表示所述权值的绝对值大于所述第二阈值，即所述权值对应的输入神经元与输出神经元之间有连接；以直接索引形式表示权值的连接关系数据有两种表示顺序：以每个输出神经元与所有输入神经元的连接状态组成一个0和1的字符串来表示所述权值的连接关系数据；或者每个输入神经元与所有输出神经元的连接状态组成一个0和1的字符串来表示所述权值的连接关系数据；

当所述权值的连接关系数据以步长索引的形式表示时，该连接关系数据为与输出神经元有连接的输入神经元的与上一个与该输出神经元有连接的输入神经元之间的距离值组成的字符串。

神经网络运算设备和方法

技术领域

[0001] 本发明涉及神经网络领域,尤其涉及一种神经网络运算设备和方法。

背景技术

[0002] 人工神经网络(Artificial Neural Networks,ANNs)简称为神经网络(Neural Networks,NNs)。它是一种模仿动物神经网络行为特征,进行分布式并行信息处理的算法数学模型。这种网络依靠系统的复杂程度,通过调整内部大量节点之间的相互连接关系,从而达到处理信息的目的。

[0003] 神经网络是一个高计算量和高访存的算法,权值越多,计算量和访存量。都会增大。为了减小计算量和权值数量,从而降低访存量,因此提出了稀疏神经网络。稀疏神经网络的定义为:在神经网络中,值为0的权值的数目很多,并且值为非0的权值分布没有规律,则称该神经网络为稀疏神经网络。定义权值为0的元素数量与神经网络所有权值元素数量的比值为神经网络的稀疏度,如图1a所示。

[0004] 现有技术中,GPU在执行图形图像运算以及科学计算时会涉及稀疏神经网络的计算,但是由于GPU是专门用来执行图形图像运算以及科学计算的设备,没有对稀疏的卷积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稀疏的人工神经网络运算,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,同时带来了巨大的功耗开销。

发明内容

[0005] 本发明实施例提供一种神经网络运算设备及方法,通过对神经网络数据进行处理,减小了在进行人工神经网络运算之前译码的工作量,进而减小了额外的开销,并且提高了运算速率。

[0006] 第一方面,本发明实施例提供了一种神经网络运算模块,包括:

[0007] 映射单元,用于接收输入数据之后,对所述输入数据进行处理,以得到处理后的输入数据,所述输入数据包括至少一个输入神经元和至少一个权值,所述处理后的输入数据包括处理后的输入神经元和处理后的权值;

[0008] 存储单元,用于存储所述处理后的输入神经元、处理后的权值、神经网络指令和运算结果;

[0009] 直接存储访问单元,用于在所述存储单元与指令缓存单元、第一输入缓存单元、第二输入缓存单元和输出缓存单元进行数据的读写;

[0010] 所述指令缓存单元,用于缓存所述直接存储访问单元读取所述神经网络指令;

[0011] 所述第一输入缓存单元,用于缓存所述直接存储访问单元读取的第一缓存数据,所述第一缓存数据为所述处理后的输入神经元或所述处理后的权值;

[0012] 所述第二输入缓存单元,用于缓存所述直接存储访问单元读取的第二缓存数据,所述第二缓存数据为所述处理后的权值或所述处理后的输入神经元,且所述第二缓存数据

与所述第二缓存数据不一致；

[0013] 指令控制单元，用于从所述指令缓存单元中获取所述神经网络指令，并将所述神经网络指令译码成运算单元执行的微指令；

[0014] 所述运算单元，用于从所述第一输入缓存单元和所述第二输入缓存单元中获取所述处理后的输入神经元和所述处理后的权值后，根据所述微指令对所述处理后的输入神经元和所述处理后的权值进行人工神经网络运算，以得到所述运算结果；

[0015] 所述输出缓存单元，用于缓存所述运算结果。

[0016] 第二方面，本发明实施例提供了另一种神经网络运算模块，包括：

[0017] 存储单元，用于存储输入数据、神经网络指令和运算结果，所述输入数据包括至少一个输入神经元和至少一个权值；

[0018] 直接内存访问直接存储访问单元，用于在所述存储单元与指令缓存单元、映射单元和输出缓存单元进行数据的读写；

[0019] 映射单元，用于通过所述直接存储访问单元获取所述输入数据后，对所述输入数据进行处理，以得到处理后的输入数据，所述处理后的输入数据包括处理后的输入神经元和处理后的权值，并将所述处理后的输入神经元和所述处理后的权值存储到第一输入缓存单元和第二输入缓存单元中；

[0020] 所述第一输入缓存单元，用于缓存第一缓存数据，所述第一缓存数据为所述处理后的输入神经元或处理后的权值；

[0021] 所述第二输入缓存单元，用于缓存第二缓存数据，所述第二缓存数据为所述处理后的输入神经元或处理后的权值，且所述第二缓存数据与所述第一缓存数据不一致；

[0022] 所述指令缓存单元，用于缓存所述直接存储访问单元读取神经网络指令；

[0023] 指令控制单元，用于从所述指令缓存单元中获取所述神经网络指令，并将所述神经网络指令译码成运算单元执行的微指令；

[0024] 所述运算单元，用于从所述第一输入缓存单元和所述第二输入缓存单元中获取所述处理后的输入神经元和所述处理后的权值后，根据所述微指令对所述处理后的输入神经元和所述处理后的权值进行人工神经网络运算，以得到所述运算结果；

[0025] 所述输出缓存单元，用于缓存所述运算结果。

[0026] 第三方面，本发明实施例提供了另一种神经网络运算模块，包括：

[0027] 存储单元，用于存储第一输入数据及所述第一输入数据的连接关系数据、处理后的第二输入数据、神经网络指令和运算结果，所述第一输入数据为输入神经元权值，所述第一输入数据的连接关系数据为输入神经元的连接关系数据或者权值的连接关系数据，所述处理后的第二输入数据为处理后的输入神经元或者处理后的权值；

[0028] 直接内存访问直接存储访问单元，用于在所述存储单元与指令缓存单元、映射单元、第一输入缓存单元和输出缓存单元进行数据的读写；

[0029] 映射单元，用于通过所述直接存储访问单元获取所述第一输入数据和所述第一输入数据的连接关系数据后，根据所述第一输入数据的连接关系数据对所述第一输入数据进行处理，以得到处理后的第一输入数据，并将所述处理后的第一输入数据存储到第一输入缓存单元中，所述处理后的第一输入数据为处理后的输入神经元或者处理后的权值；

[0030] 所述第一输入缓存单元，用于缓存所述处理后的第一输入数据；

[0031] 所述第二输入缓存单元，用于缓存所述处理后的第二输入数据，且所述处理后的第一输入数据与所述处理后的第二输入数据不一致；

[0032] 所述指令缓存单元，用于缓存所述直接存储访问单元读取神经网络指令；

[0033] 指令控制单元，用于从所述指令缓存单元中获取所述神经网络指令，并将所述神经网络指令译码成运算单元执行的微指令；

[0034] 所述运算单元，用于从所述第一输入缓存单元和所述第二输入缓存单元中获取所述处理后的第一输入数据和所述处理后的第二输入数据后，根据所述微指令对所述处理后的第一输入数据和所述处理后的第二输入数据进行人工神经网络运算，以得到所述运算结果；

[0035] 所述输出缓存单元，用于缓存所述运算结果。

[0036] 第四方面，本发明实施例提供了一种神经网络运算方法，包括：

[0037] 对输入数据进行处理，以得到处理后的输入数据；

[0038] 获取神经运算指令，将所述神经运算指令译码成微指令；

[0039] 根据所述微指令对所述处理后的输入数据进行人工神经网络运算，以得到运算结果。

[0040] 第五方面，本发明实施例提供了一种神经网络运算装置，该神经网络运算装置包括一个或者多个第一方面所述的神经网络运算模块、第二方面所述的神经网络运算模块或者第三方面所述的神经网络运算模块。该神经网络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；

[0041] 当所述神经网络运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；

[0042] 其中，多个所述神经网络运算模块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；多个所述神经网络运算模块共享同一控制系统或拥有各自的控制系统；多个所述神经网络运算模块共享内存或者拥有各自的内存；多个所述神经网络运算模块的互联方式是任意互联拓扑。

[0043] 第六方面，本发明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五方面所述的神经网络处理装置、通用互联接口，和其他处理装置。该神经网络运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。

[0044] 第七方面，本发明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神经网络运算模块、上述第二方面所述的神经网络运算模块、上述第三方面所述的神经网络运算模块、上述第五方面所述的神经网络运算装置或者上述第六方面所述的组合处理装置。

[0045] 第八方面，本发明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七方面所述的神经网络芯片；

[0046] 第九方面，本发明实施例提供了一种板卡，该板卡包括上述第八方面所述的神经网络芯片封装结构。

[0047] 第十方面，本发明实施例提供了一种电子装置，该电子装置包括上述第七方面所

述的神经网络芯片或者上述第九方面所述的板卡。

[0048] 可以看出，在本发明实施例的方案中，映射单元对输入神经元和权值进行处理，以得到处理后的输入神经元和处理后的权值，运算单元根据指令控制单元对神经网络指令进行译码得到的微指令对处理后的输入神经元和处理后的权值进行人工神经网络运算。与现有技术相比，采用本发明实施例减小了在进行人工神经网络运算之前译码的工作量，进而减小了额外的开销，并且提高了运算速率。

[0049] 本发明的这些方面或其他方面在以下实施例的描述中会更加简明易懂。

附图说明

[0050] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，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。

[0051] 图1a为本发明实施例提供的一种稀疏神经网络结构示意图；

[0052] 图1b为本发明实施例提供的一种神经网络运算模块的结构示意图；

[0053] 图2为本发明实施例提供的一种神经网络运算模块的局部结构示意图；

[0054] 图3为本发明实施例提供的一种神经网络结构示意图；

[0055] 图4为本发明实施例提供的另一种神经网络运算模块的局部结构示意图；

[0056] 图5为本发明实施例提供的另一种神经网络运算模块的结构示意图；

[0057] 图6为本发明实施例提供的另一种神经网络运算模块的结构示意图；

[0058] 图7为本发明实施例提供的另一种神经网络运算模块的局部结构示意图；

[0059] 图8为本发明实施例提供的另一种神经网络运算模块的局部结构示意图；

[0060] 图9为本发明实施例提供的另一种神经网络运算模块的局部结构示意图；

[0061] 图10为本发明实施例提供的另一种神经网络运算模块的局部结构示意图；

[0062] 图11为本发明实施例提供的另一种神经网络运算模块的局部结构示意图；

[0063] 图12为本发明实施例提供的另一种神经网络结构示意图；

[0064] 图13为本发明实施例提供的另一种神经网络结构示意图；

[0065] 图14为本发明实施例提供的另一种神经网络结构示意图；

[0066] 图15为本发明实施例提供的另一种神经网络结构示意图；

[0067] 图16a为本发明实施例提供的一种组合处理装置的结构示意图；

[0068] 图16b为本发明实施例提供的另一种组合处理装置的结构示意图；

[0069] 图17为本发明实施例提供的一种板卡的结构示意图；

[0070] 图18为本发明实施例提供的一种神经网络芯片封装结构的示意图；

[0071] 图19为本发明实施例提供的另一种神经网络芯片封装结构的示意图；

[0072] 图20为本发明实施例提供的另一种神经网络芯片封装结构的示意图；

[0073] 图21为本发明实施例提供的一种神经网络运算方法的流程示意图。

具体实施方式

[0074] 以下分别进行详细说明。

[0075] 本发明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。

[0076] 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发明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。

[0077] 参见图1b，图1b为本发明实施例提供的一种神经网络运算模块的结构示意图。该神经网络运算模块用于加速稀疏神经网络的运算。如图1b所示，该神经网络运算模块100包括：映射单元101、存储单元102、直接存储访问(Direct Memory Access,DMA)单元103、指令缓存单元104、第一输入缓存单元105、第二输入缓存单元106、指令控制单元107、运算单元108和输出缓存单元109。

[0078] 其中，上述映射单元101，用于接收输入数据之后，对所述输入数据进行处理，以得到处理后的输入数据，所述输入数据包括至少一个输入神经元和至少一个权值，所述处理后的输入数据包括处理后的输入神经元和处理后的权值。

[0079] 上述输入数据包括至少一个输入神经元和至少一个权值。上述映射单元101确定所述至少一个输入神经元中每个输入神经元的绝对值是否大于第一阈值。当上述输入神经元的绝对值小于或者等于该第一阈值时，上述映射单元101将该输入神经元删除；当上述输入神经元的绝对值大于上述第一阈值时，上述映射单元101保留该输入神经元，该映射单元101将删除后的输出神经元输出，作为处理后的输入神经元。上述映射单元101获取输入神经元的连接关系数据，该输入神经元的连接关系数据表示上述至少一个输入神经元中绝对值大于上述第一阈值的输入神经元的位置信息。上述映射单元101确定上述至少一个权值中每个权值的绝对值是否大于第二阈值。当权值的绝对值小于或者等于上述第二阈值时，上述映射单元101将该权值删除，并根据上述输入神经元的连接关系数据将从上述删除后的权值中选择相关的权值输出，作为处理后的权值。

[0080] 在一种可行的实施例中，上述输入数据包括至少一个输入神经元和至少一个权值。上述映射单元101确定所述至少一个权值中每个权值的绝对值是否大于第二阈值。当上述权值的绝对值小于或者等于该第二阈值时，上述映射单元101将该权值删除；当上述权值的绝对值大于上述第二阈值时，上述映射单元101保留该权值，该映射单元101将删除后的权值输出，作为处理后的权值。上述映射单元101获取权值的连接关系数据，该权值的连接关系数据表示上述至少一个输入神经元与输出神经元之间的连接关系的数据。上述映射单元101确定上述至少一个输入神经元中每个输入神经元的绝对值是否大于第一阈值。当输入神经元的绝对值小于或者等于上述第一阈值时，上述映射单元101将该输入神经元删除，并根据上述权值的连接关系数据将从上述删除后的输入神经元中选择相关的输入神经元输出，作为处理后的输入神经元。

[0081] 进一步地，上述映射单元101将上述处理后的输入神经元和处理后的权值按照一一对应的格式存储到上述存储单元102中。

[0082] 具体地，上述映射单元101对上述处理后的输入神经元和上述处理后的权值按照一一对应的格式进行存储的具体方式是将上述处理后的输入神经元中的每个处理后的输入神经元和与其对应的处理后的权值作为一个数据集，并将该数据集存储到上述存储单元102中。

[0083] 具体地，如图2所示，上述映射单元101包括：

[0084] 第一稀疏处理单元1011，用于对第二输入数据进行处理，以得到第三输出数据和第二输出数据，并将所述第三输出数据传输至第一数据处理单元1012。

[0085] 第一数据处理单元1012，用于接收第一输入数据和接收所述第三输出数据，并根据上述第三输出数据和第一输入数据输出第一输出数据。

[0086] 其中，当所述第一输入数据包括至少一个输入神经元，所述第二输入数据包括至少一个权值时，所述第一输出数据为处理后的输入神经元，所述第二输出数据为处理后的权值，所述第三输出数据为权值的连接关系数据；当所述第一输入数据包括至少一个权值，所述第二输入数据包括至少一个输入神经元时，所述第一输出数据为处理后的权值，所述第二输出数据为处理后的输入神经元，所述第三输出数据为输入神经元的连接关系数据。

[0087] 具体地，当上述第二输入数据为权值时，且权值的形式为w_{ij}，该w_{ij}表示第i个输入神经元与第j个输出神经元之间的权值；上述第一稀疏处理单元1011根据权值确定上述连接关系数据(即上述第三输出数据)，并将上述权值中绝对值小于或者等于第二阈值的权值删除，得到处理后的权值(即上述第二输出数据)；当上述第二输入数据为输入神经元时，上述第一稀疏处理单元1011根据输入神经元得到连接关系数据，并将该输入神经元中的绝对值小于或等于上述第一阈值的输入神经元删除，以得到处理后的输入神经元。

[0088] 可选地，上述第一阈值可为0.1、0.08、0.05、0.02、0.01、0或者其他值。

[0089] 可选地，上述第二阈值可为0.1、0.08、0.06、0.05、0.02、0.01、0或者其他值。

[0090] 需要指出的是，上述第一阈值和上述第二阈值可以一致，也可以不一致。

[0091] 其中，上述连接关系数据可以步长索引或者直接索引的形式表示。

[0092] 具体地，以直接索引形式表示的连接关系数据为由0和1组成的字符串，当上述第二输入数据为权值时，0表示该权值的绝对值小于或者等于上述第二阈值，即该权值对应的输入神经元与输出神经元之间没有连接，1表示该权值的绝对值大于上述第二阈值，即该权值对应的输入神经元与输出神经元之间有连接。以直接索引形式表示的连接关系数据有两种表示顺序：以每个输出神经元与所有输入神经元的连接状态组成一个0和1的字符串来表示权值的连接关系；或者每个输入神经元与所有输出神经元的连接状态组成一个0和1的字符串来表示权值的连接关系。当上述第二输入数据为输入神经元时，0表示该输入神经元的绝对值小于或者等于上述第一阈值，1表示该输入神经元的绝对值大于上述第一阈值。

[0093] 当上述第二输入数据为权值时，以步长索引形式表示的连接关系数据为与输出神经元有连接的输入神经元与上一个与该输出神经元有连接的输入神经元之间的距离值组成的字符串；当上述第二输入数据为输入神经元时，以步长索引表示的数据以当前绝对值大于上述第一阈值的输入神经元与上一个绝对值大于上述第一阈值的输入神经元之间的距离值组成的字符串表示。

[0094] 举例说明，假设上述第一阈值和上述第二阈值均为为0.01，参见图3，图3为本发明实施例提供的一种神经网络的示意图。如图3中的a图所示，上述第一输入数据为输入神经

元，包括输入神经元i1、i2、i3和i4，上述第二输入数据为权值。对于输出神经元o1，权值为w_{11}，w_{21}，w_{31}和w_{41}；对于输出神经元o2，权值w_{12}，w_{22}，w_{32}和w_{42}，其中权值w_{21}，w_{12}和w_{42}的值为0，其绝对值均小于上述第一阈值0.01，上述第一稀疏处理单元1011确定上述输入神经元i2和输出神经元o1没有连接，上述输入神经元i1和i4与输出神经元o2没有连接，上述输入神经元i1、i3和i4与上述输出神经元o1有连接，上述输入神经元i2和i3与输出神经元o2有连接。以每个输出神经元与所有输入神经元的连接状态表示上述连接关系数据，则上述输出神经元o1的连接关系数据为“1011”，输出神经元o2的连接关系数据为“0110”(即上述连接关系数据为“10110110”)；以每个输入神经元与所有输出神经元的连接关系，则输入神经元i1的连接关系数据为“10”，输入神经元i2的连接关系数据为“01”，输入神经元i3的连接关系数据为“11”，输入神经元i4的连接关系数据为“10”(即上述连接关系数据为“10011110”)。

[0095] 对于上述输出神经元o1，上述映射单元101将上述i1与w_{11}，i3与w_{31}和i4与w_{41}分别作为一个数据集，并将该数据集存储到上述存储单元102中；对于输出神经元o2，上述映射单元101将上述i2与w_{22}和i3与w_{32}分别作为一个数据集，并将该数据集存储到上述存储单元102中。

[0096] 针对上述输出神经元o1，上述第二输出数据为w_{11}，w_{31}和w_{41}；针对上述输出神经元o2，上述第二输出数据为w_{22}和w_{32}。

[0097] 当上述第二输入数据为输入神经元i1、i2、i3和i4，且该输入神经元的值分别为1，0，3，5则上述连接关系数据(即上述第三输出数据)为“1011”，上述第二输出数据为1，3，5。

[0098] 如图3中的b图所示，上述第一输入数据包括输入神经元i1、i2、i3和i4，上述第二输入数据为权值。对于输出神经元o1，权值为w_{11}，w_{21}，w_{31}和w_{41}；对于输出神经元o2，权值w_{12}，w_{22}，w_{32}和w_{42}，其中权值w_{21}，w_{12}和w_{42}的值为0，上述稀疏处理单元1011确定上述输入神经元i1、i3和i4与上述输出神经元o1有连接，上述输入神经元i2和i3与输出神经元o2有连接。上述输出神经元o1与输入神经元之间的连接关系数据为“021”。其中，该连接关系数据中第一个数字“0”表示第一个与输出神经元o1有连接的输入神经元与第一个输入神经元之间的距离为0，即第一个与输出神经元o1有连接的输入神经元为输入神经元i1；上述连接关系数据中第二个数字“2”表示第二个与输出神经元o1有连接的输入神经元与第一个与输出神经元o1有连接的输入神经元(即输入神经元i1)之间的距离为2，即第二个与输出神经元o1有连接的输入神经元为输入神经元i3；上述连接关系数据中第三个数字“1”表示第三个与输出神经元o1有连接的输入神经元与第二个与该输出神经元o1有连接的输入神经元之间的距离为1，即第三个与输出神经元o1有连接的输入神经元为输入神经元i4。

[0099] 上述输出神经元o2与输入神经元之间的连接关系数据为“11”。其中，该连接关系数据中的第一数字“1”表示第一个与输出神经元o2有连接的输入神经元与第一个输入神经元(即输入神经元i1)之间的距离为，即该第一个与输出神经元o2有连接关系的输入神经元为输出神经元i2；上述连接关系数据中的第二数字“1”表示第二个与输出神经元o2有连接的输入神经元与第一个与输出神经元o2有连接的输入神经元的距离为1，即第二个与输出神经元o2有连接的输入神经元为输入神经元i3。

[0100] 对于上述输出神经元o1，上述映射单元101将上述i1与w_{11}，i3与w_{31}和i4与w_{41}分别作为一个数据集，并将该数据集存储到上述存储单元102中；对于输出神经元o2，上述映射单元101将上述i2与w_{22}和i3与w_{32}分别作为一个数据集，并将该数据集存储到上述存储单元

102中。

[0101] 针对上述输出神经元o1，上述第二输出数据为w_{11}，w_{31}和w_{41}；针对上述输出神经元o2，上述第二输出数据为w_{22}和w_{32}。

[0102] 当上述第二输入数据为输入神经元i1、i2、i3和i4，且该输入神经元的值分别为1，0，3，5则上述连接关系数据即上述第三输出数据为“021”，上述第二输出数据为1，3，5。

[0103] 当上述第一输入数据为输入神经元时，则上述第二输入数据为权值，上述第三输出数据为输出神经元与上述输入神经元之间的连接关系数据。上述第一数据处理单元1012接收到上述输入神经元后，将该输入神经元中绝对值小于或等于上述第二阈值的输入神经元剔除，并根据上述连接关系数据，从剔除后的输入神经元中选择与上述权值相关的输入神经元，作为第一输出数据输出。

[0104] 举例说明，假设上述第一阈值为0，上述输入神经元i1、i2、i3和i4，其值分别为1，0，3和5，对于输出神经元o1，上述第三输出数据(即连接关系数据)为“021”，上述第二输出数据为w_{11}，w_{31}和w_{41}。上述第一数据处理单元1012将上述输入神经元i1、i2、i3和i4中值为0的输入神经元剔除，得到输入神经元i1、i3和i4。该第一数据处理单元1012根据上述第三输出数据“021”确定上述输入神经元i1、i3和i4均与上述输出神经元均有连接，故上述数据处理单元1012将上述输入神经元i1、i3和i4作为第一输出数据输出，即输出1，3，5。

[0105] 当上述第一输入数据为权值，上述第二输入数据为输入神经元时，上述第三输出数据为上述输入神经元的连接关系数据。上述第一数据处理单元1012接收到上述权值w_{11}，w_{21}，w_{31}和w_{41}后，将该权值中绝对值小于上述第一阈值的权值剔除，并根据上述连接关系数据，从上述剔除后的权值中选择与该上述输入神经元相关的权值，作为第一输出数据并输出。

[0106] 举例说明，假设上述第二阈值为0，上述权值w_{11}，w_{21}，w_{31}和w_{41}，其值分别为1，0，3和4，对于输出神经元o1，上述第三输出数据(即连接关系数据)为“1011”，上述第二输出数据为i1，i3和i5。上述第一数据处理单元1012将上述权值w_{11}，w_{21}，w_{31}和w_{41}中值为0的输入神经元剔除，得到权值w_{11}，w_{21}，w_{31}和w_{41}。该第一数据处理单元1012根据上述第三输出数据“1011”确定上述输入神经元i1、i2，i3和i4中的输入神经元i2的值为0，故上述第一数据处理单元1012将上述输入神经元1，3和4作为第一输出数据输出。

[0107] 在一种可行的实施例中，第三输入数据和第四输入数据分别为至少一个权值和至少一个输入神经元，上述映射单元101确定上述至少一个输入神经元中绝对值大于上述第一阈值的输入神经元的位置，并获取输入神经元的连接关系数据；上述映射单元101确定上述至少一个权值中绝对值大于上述第二阈值的权值的位置，并获取权值的连接关系数据。上述映射单元101根据上述权值的连接关系数据和输入神经元的连接关系数据得到一个新的连接关系数据，该连接关系数据表示上述至少一个输入神经元中绝对值大于上述第一阈值的输入神经元与输出神经元之间的关系和对应的权值的值。101映射单元101根据该新的连接关系数据、上述至少一个输入神经元和上述至少一个权值获取处理后的输入神经元和处理后的权值。

[0108] 进一步地，上述映射单元101将上述处理后的输入神经元和处理后的权值按照一一对应的格式存储到上述存储单元102中。

[0109] 具体地，上述映射单元101对上述处理后的输入神经元和上述处理后的权值按照

一一对应的格式进行存储的具体方式是将上述处理后的输入神经元中的每个处理后的输入神经元和与其对应的处理后的权值作为一个数据集，并将该数据集存储到上述存储单元102中。

[0110] 对于映射单元101包括第一稀疏处理单元1011和第一数据处理单元1012的情况，映射单元101中的稀疏处理单元1011对输入神经元或者权值进行稀疏化处理，减小了权值或者输入神经元的数量，进而减小了运算单元进行运算的次数，提高了运算效率。

[0111] 具体地，如图4所示，上述映射单元101包括：

[0112] 第二稀疏处理单元1013，用于接收到第三输入数据后，根据所述第三输入数据得到第一连接关系数据，并将该第一连接关系数据传输至连接关系处理单元1015；

[0113] 第三稀疏处理单元1014，用于接收到第四输入数据后，根据所述第四输入数据得到第二连接关系数据，并将该第二连接关系数据传输至所述连接关系处理单元1015；

[0114] 所述连接关系处理单元1015，用于根据所述第一连接关系数据和所述第二连接关系数据，以得到第三连接关系数据，并将该第三连接关系数据传输至第二数据处理单元1016；

[0115] 所述第二数据处理单元1016，用于在接收到所述第三输入数据，所述第四输入数据和所述第三连接关系数据后，根据所述第三连接关系数据对所述第三输入数据和所述第四输入数据进行处理，以得到第四输出数据和第五输出数据；

[0116] 其中，当所述第三输入数据包括至少一个输入神经元，第四输入数据包括至少一个权值时，所述第一连接关系数据为输入神经元的连接关系数据，所述第二连接关系数据为权值的连接关系数据，所述第四输出数据为处理后的输入神经元，所述第五输出数据为处理后的权值；当所述第三输入数据包括至少一个权值，所述第四输入数据包括至少一个输入神经元时，所述第一连接关系数据为权值的连接关系数据，所述第二连接关系数据为输入神经元的连接关系数据，所述第四输出数据为处理后的权值，所述第五输出数据为处理后的输入神经元。

[0117] 当上述第三输入数据包括至少一个输入神经元时，上述第一连接关系数据为用于表示该至少一个输入神经元中绝对值大于上述第一阈值的输入神经元的位置的字符串；当上述第三输入数据包括至少一个权值时，上述第一连接关系数据为用于表示输入神经元与输出神经元之间是否有连接的字符串。

[0118] 当上述第四输入数据包括至少一个输入神经元时，上述第二连接关系数据为用于表示该至少一个输入神经元中绝对值大于上述第一阈值的输入神经元的位置的字符串；当上述第四输入数据包括至少一个权值时，上述第二连接关系数据为用于表示输入神经元与输出神经元之间是否有连接的字符串。

[0119] 需要说明的是，上述第一连接关系数据、第二连接关系数据和第三连接关系数据均可以步长索引或者直接索引的形式表示，具体可参见上述相关描述。

[0120] 换句话说，上述连接关系处理单元1015对上述第一连接关系数据和上述第二连接关系数据进行处理，以得到第三连接关系数据。该第三连接关系数据可以直接索引或者步长索引的形式表示。

[0121] 具体地，上述当上述第一连接关系数据和上述第二连接关系数据均以直接索引的形式表示时，上述连接关系处理单元1015对上述第一连接关系数据和上述第二连接关系数

据进行与操作，以得到第三连接关系数据，该第三连接关系数据是以直接索引的形式表示的。

[0122] 需要说明的是，表示上述第一连接关系数据和第二连接关系数据的字符串在内存中是按照物理地址高低的顺序存储的，可以是由高到低的顺序存储的，也可以是由低到高的顺序存储的。

[0123] 当上述第一连接关系数据和上述第二连接关系数据均以步长索引的形式表示，且表示上述第一连接关系数据和第二连接关系数据的字符串是按照物理地址由低到高的顺序存储时，上述连接关系处理单元1015将上述第一连接关系数据的字符串中的每一个元素与存储物理地址低于该元素存储的物理地址的元素进行累加，得到的新的元素组成第四连接关系数据；同理，上述连接关系处理单元1015对上述第二连接关系数据的字符串进行同样的处理，得到第五连接关系数据。然后上述连接关系处理单元1015从上述第四连接关系数据的字符串和上述第五连接关系数据的字符串中，选取相同的元素，按照元素值从小到大的顺序排序，组成一个新的字符串。上述连接关系处理单元1015将上述新的字符串中将每一个元素与其相邻且值小于该元素值的元素进行相减，以得到一个新的元素。按照该方法，对上述新的字串中的每个元素进行相应的操作，以得到上述第三连接关系数据。

[0124] 举例说明，假设以步长索引的形式表示上述第一连接关系数据和上述第二连接关系数据，上述第一连接关系数据的字符串为“01111”，上述第二连接关系数据的字符串为“022”，上述连接关系处理单元1015将上述第一连接关系数据的字符串中的每个元素与其相邻的前一个元素相加，得到第四连接关系数据“01234”；同理，上述连接关系处理单元1015对上述第二连接关系数据的字符串进行相同的处理后得到的第五连接关系数据为“024”。上述连接关系处理单元1015从上述第四连接关系数据“01234”和上述第五连接关系数据“024”选组相同的元素，以得到新的字符串“024”。上述连接关系处理单元1015将该新的字符串中的每个元素与其相邻的前一个元素进行相减，即0，(2-0)，(4-2)，以得到上述第三连接数据“022”。

[0125] 当上述第一连接关系数据和上述第二连接关系数据中的任意一个以步长索引的形式表示，另一个以直接索引的形式表示时，上述连接关系处理单元1015将上述以步长索引表示的连接关系数据转换成以直接索引的表示形式或者将以直接索引表示的连接关系数据转换成以步长索引表示的形式。然后上述连接关系处理单元1015按照上述方法进行处理，以得到上述第三连接关系数据(即上述第五输出数据)。

[0126] 可选地，当上述第一连接关系数据和上述第二连接关系数据均以直接索引的形式表示时，上述连接关系处理单元1015将上述第一连接关系数据和上述第二连接关系数据均转换成以步长索引的形式表示的连接关系数据，然后按照上述方法对上述第一连接关系数据和上述第二连接关系数据进行处理，以得到上述第三连接关系数据。

[0127] 具体地，上述第三输入数据可为输入神经元或者权值、第四输入数据可为输入神经元或者权值，且上述第三输入数据和第四输入数据不一致。上述第二数据处理单元1016根据上述第三连接关系数据从上述第三输入数据(即输入神经元或者权值)中选取与该第三连接关系数据相关的数据，作为第四输出数据；上述第二数据处理单元1016根据上述第三连接关系数据从上述第四输入数据中选取与该第三连接关系数据相关的数据，作为第五输出数据。

[0128] 进一步地，上述第二数据处理单元1016将上述处理后的输入神经元中的每个处理后的输入神经元与其对应的处理后的权值作为一个数据集，将该数据集存储出上述存储单元102中。

[0129] 举例说明，假设上述第三输入数据包括输入神经元i1，i2，i3和i4，上述第四输入数据包括权值w_{11}，w_{21}，w_{31}和w_{41}，上述第三连接关系数据以直接索引方式表示，为“1010”，则上述第二数据处理单元1016输出的第四输出数据为输入神经元i1和i3，输出的第五输出数据为权值w_{11}和w_{31}。上述第二数据处理单元1016将输入神经元i1与权值w_{11}和输入神经元i3与权值w_{31}分别作为一个数据集，并将该数据集存储到上述存储单元102中。

[0130] 对于映射单元101包括第二稀疏处理单元1013，第三稀疏处理单元1014、连接关系处理单元1015和第二数据处理单元1016的情况，映射单元101中的稀疏处理单元对输入神经元和权值均进行稀疏化处理，使得输入神经元和权值的数量进一步减小，进而减小了运算单元的运算量，提高了运算效率。

[0131] 可选地，所述映射单元101对所述输入数据进行处理之前，所述映射单元101还用于：

[0132] 对所述至少一个输入神经元进行分组，以得到M组输入神经元，所述M为大于或者等于1的整数；

[0133] 判断所述M组输入神经元的每一组输入神经元是否满足第一预设条件，所述第一预设条件包括一组输入神经元中绝对值小于或者等于第三阈值的输入神经元的个数小于或者等于第四阈值；

[0134] 当所述M组输入神经元任意一组输入神经元不满足所述第一预设条件时，将该组输入神经元删除；

[0135] 对所述至少一个权值进行分组，以得到N组权值，所述N为大于或者等于1的整数；

[0136] 判断所述N组权值的每一组权值是否满足第二预设条件，所述第二预设条件包括一组权值中绝对值小于或者等于第五阈值的权值的个数小于或者等于第六阈值；

[0137] 当所述N组权值任意一组权值不满足所述第二预设条件时，将该组权值删除。

[0138] 可选地，上述第三阈值可为0.5，0.2，0.1，0.05，0.025，0.0，0或者其他值。

[0139] 其中，上述第四阈值与上述一组输入神经元中输入神经元的个数相关。可选地，该第四阈值＝一组输入神经元中的输入神经元个数-1或者该第四阈值为其他值。

[0140] 可选地，上述第五阈值可为0.5，0.2，0.1，0.05，0.025，0.01，0或者其他值。

[0141] 其中，上述第六阈值与上述一组权值中的权值个数相关。可选地，该第六阈值＝一组权值中的权值个数-1或者该第六阈值为其他值。

[0142] 需要说明的是，上述第三阈值和上述第五阈值可相同或者不同，上述第四阈值和上述第六阈值可相同或者不同。

[0143] 上述存储单元102，用于存储上述处理后的输入神经元、处理后的权值和神经网络指令。

[0144] 上述直接存储访问单元103，用于在上述存储单元102与上述指令缓存单元104、第一输入缓存单元105、第二输入缓存单元106和输出缓存单元109之间进行数据的读写。

[0145] 具体地，上述直接存储访问单元103从上述存储单元102中读取神经网络指令，并将该神经网络指令写入上述指令缓存单元104中。上述直接存储访问单元103从上述存储单

元102读取上述处理后的输入神经元和处理后的权值，并将该输入神经元和权值分别写入上述第一输入缓存单元105和上述第二输入缓存单元106，或者分别写入上述第二输入缓存单元106和上述第一输入缓存单元105。

[0146] 上述指令缓存单元104，用于缓存上述直接存储访问单元103读取的神经网络指令。

[0147] 上述第一输入缓存单元105，用于缓存上述直接存储访问单元103读取的处理后的输入神经元或处理后的权值。

[0148] 上述第二输入缓存单元106，用于缓存上述直接存储访问单元103读取的处理后的输入神经元或处理后的权值。

[0149] 需要说明的是，当上述第一输入缓存单元105用于缓存处理后的输入神经元时，则上述第二输入缓存单元106用于缓存处理后的权值；当上述第二输入缓存单元105用于缓存处理后的权值时，则上述第一输入缓存单元106用于缓存处理后的输入神经元。

[0150] 需要说明的是，第一阈值、第二阈值、第三阈值、第四阈值、第五阈值和第六阈值可均存储在上述存储单元102、第一输出缓存单元105或者第二输入缓存单元106中；上述第一阈值、第二阈值、第三阈值、第四阈值、第五阈值和第六阈值中部分阈值存储在上述存储单元102、部分阈值存储在上述第一输出缓存单元105中，部分阈值存储在上述第二输出缓存单元106中。

[0151] 上述指令控制单元107，用于从上述指令缓存单元104中获取神经网络指令，并将该神经网络指令译码成上述运算单元108执行的微指令。

[0152] 上述运算单元108，在从上述第一输入缓存单元105和上述第二输入缓存单元106中获取上述处理后的输入神经元和处理后的权值，根据上述微指令对上述处理后的权值和处理后的输入神经元进行人工神经网络运算，得到运算结果，并将该运算结果存储到上述输出缓存单元109中，该输出缓存单元109通过上述直接存储访问单元103将该运算结果存储到上述存储单元102中。

[0153] 需要指出的是，上述指令缓存单元104、上述第一输入缓存单元105、上述第二输入缓存单元106和上述输出缓存单元109均可为片上缓存。

[0154] 进一步地，上述运算单元108包括但不限定于三个部分，分别为乘法器、一个或多个加法器(可选地，多个加法器组成加法树)和激活函数单元/激活函数运算器。上述乘法器将第一输入数据(in1)和第二输入数据(in2)相乘得到第一输出数据(out1)，过程为：out1＝in1*in2；上述加法树将第三输入数据(in3)通过加法树逐级相加得到第二输出数据(out2)，其中in3是一个长度为N的向量，N大于1，过称为：out2＝in3[1]+in3[2]+...+in3[N]，和/或将第三输入数据(in3)通过加法树累加之后得到的结果和第四输入数据(in4)相加得到第二输出数据(out2)，过程为：out2＝in3[1]+in3[2]+...+in3[N]+in4，或者将第三输入数据(in3)和第四输入数据(in4)相加得到第二输出数据(out2)，过称为：out2＝in3+in4；上述激活函数单元将第五输入数据(in5)通过激活函数(active)运算得到第三输出数据(out3)，过程为：out3＝active(in5)，激活函数active可以是sigmoid、tanh、relu、softmax等函数，除了做激活操作，激活函数单元可以实现其他的非线性函数运算，可将输入数据(in)通过函数(f)运算得到输出数据(out)，过程为：out＝f(in)。

[0155] 上述运算单元108还可以包括池化单元，池化单元将输入数据(in)通过池化运算

得到池化操作之后的输出数据(out)，过程为out＝pool(in)，其中pool为池化操作，池化操作包括但不限于：平均值池化，最大值池化，中值池化，输入数据in是和输出out相关的一个池化核中的数据。

[0156] 可以看出，在本发明实施例的方案中，上述映射单元中的稀疏处理单元对输入神经元和权值进行处理，剔除绝对值小于或等于上述阈值的输入神经元和权值，减少了输入神经元和权值的数量，减少了额外的开销，运算单元根据处理后的输入神经元和权值进行人工神经网络运算，提高了运算的效率。

[0157] 需要说明的是，上述神经网络运算模块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。上述神经网络运算模块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。

[0158] 参见图5，图5为本发明实施例提供的另一种神经网络运算模块的结构示意图，如图5所示，该神经网络运算模块包括存储单元502、直接存储访问单元503、映射单元501、指令缓存单元504、第一输入缓存单元505、第二输入缓存单元506、指令控制单元507、运算单元508和输出缓存单元509。

[0159] 上述存储单元502，用于存储输入数据、神经网络指令和运算结果，所述输入数据包括至少一个输入神经元和至少一个权值。

[0160] 上述直接存储访问单元503，用于在所述存储单元502与指令缓存单元504、映射单元501和输出缓存单元509之间进行数据的读写。

[0161] 具体地，上述直接存储访问单元从上述存储单元502中读取神经网络指令，并将该神经网络指令写入上述指令缓存单元504中。上述直接存储访问单元503从上述存储单元502读取上述输入神经元和权值，并将该输入神经元和权值写入上述映射单元501。上述直接存储访问单元503从上述输出缓存单元509读取上述运算结果后，将该运算结果写入上述存储单元502中。

[0162] 上述映射单元501，用于通过所述直接存储访问单元503获取所述输入数据后，对所述输入数据进行处理，以得到处理后的输入数据，所述处理后的输入数据包括处理后的输入神经元和处理后的权值，并将所述处理后的输入神经元和所述处理后的权值存储到第一输入缓存单元505和第二输入缓存单元506中。

[0163] 具体的，上述映射单元501将上述处理后的输入神经元和上述处理后的权值分别存储到上述第一输入缓存单元505和第二输入缓存单元506中，或者分别存储到上述第二输入缓存单元506和第一输入缓存单元505。

[0164] 需要说明的是，上述映射单元501的具体功能可参见上述图1所示的实施例中映射单元101(包括第二稀疏处理单元1013、第三稀疏处理单元1014、连接关系处理单元1015和第二数据处理单元1016)的相关描述，在此不再叙述。在本实施例中第一输入数据和第二输入数据分别与图1所示实施例中的第三输入数据和第四输入数据一致，本实施例中的第一输出数据和第二输出数据分别与图1所示实施例中的第四输出数据和第五输出数据一致。

[0165] 上述第一输入缓存单元505，用于缓存第一缓存数据，所述第一缓存数据为所述处理后的输入神经元或处理后的权值。

[0166] 上述第二输入缓存单元506，用于缓存第二缓存数据，所述第二缓存数据为所述处

理后的输入神经元或处理后的权值，且所述第二缓存数据与所述第一缓存数据不一致。

[0167] 上述指令缓存单元504，用于缓存所述直接存储访问单元503读取神经网络指令。

[0168] 上述指令控制单元507，用于从所述指令缓存单元504中获取所述神经网络指令，并将所述神经网络指令译码成运算单元508执行的微指令。

[0169] 上述运算单元508，用于从所述第一输入缓存单元505和所述第二输入缓存单元506中获取所述处理后的输入神经元和所述处理后的权值后，根据所述微指令对所述处理后的输入神经元和所述处理后的权值进行人工神经网络运算，以得到所述运算结果。

[0170] 需要说明的是，上述运算单元508的功能描述可参见上述图1所示的运算单元108的相关描述，在此不再叙述。

[0171] 上述输出缓存单元509，用于缓存所述运算结果。

[0172] 需要说明的是，第一阈值、第二阈值、第三阈值、第四阈值、第五阈值和第六阈值可均存储在上述存储单元502、第一输出缓存505或者第二输入缓存506中；上述第一阈值、第二阈值、第三阈值、第四阈值、第五阈值和第六阈值中部分阈值存储在上述存储单元502、部分阈值存储在上述第一输出缓存505中，部分阈值存储在上述第二输出缓存506中。

[0173] 需要指出的是，上述指令缓存单元504、上述第一输入缓存单元505、上述第二输入缓存单元506和上述输出缓存单元509均可为片上缓存。

[0174] 可以看出，在本发明实施例的方案中，映射单元中的第四稀疏处理单元和第五稀疏处理单元分别对输入神经元和权值进行处理，以分别得到第一连接关系数据和第二连接关系数据。第二连接关系处理单元对上述第一连接关系数据和第二连接关系数据进行处理，得到第三连接关系数据。第三数据处理单元根据第三连接关系数据对上述第一输入数据进行处理以得到第一输出数据并输出，同理，根据第三连接关系数据对上述第二输入数据进行处理以得到第二输出数据并输出。通过对输入神经元和权值进行处理，得到处理后的输入神经元和权值，减小了输入数据的数据量，进而减小了总的运算量，提高了运算速度，同时也减小了额外的开销。

[0175] 需要说明的是，上述神经网络运算模块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。上述神经网络运算模块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。

[0176] 参见图6、图6为本发明实施例提供的另一种神经网络运算模块的结构示意图。如图6所示，该神经网络运算模块包括存储单元602、直接存储访问单元603、映射单元601、指令缓存单元604、第一输入缓存单元605、第二输入缓存单元606、指令控制单元607、运算单元608和输出缓存单元609。

[0177] 其中，上述存储单元602，用于存储第一输入数据及所述第一输入数据的连接关系数据、处理后的第二输入数据、神经网络指令和运算结果，所述第一输入数据为输入神经元权值，所述第一输入数据的连接关系数据为输入神经元的连接关系数据或者权值的连接关系数据，所述处理后的第二输入数据为处理后的输入神经元或者处理后的权值。

[0178] 上述直接存储访问单元603，用于在上述存储单元602与上述指令缓存单元604、上述映射单元601、上述第一输入缓存单元605和上述输出缓存单元609之间进行数据读写。

[0179] 具体地，上述直接存储访问单元603从上述存储单元602中读取上述神经网络指

令，并将该神经网络指令写入上述指令缓存单元604中；

[0180] 从上述存储单元602中读取上述输入神经元及该输入神经元的连接关系数据，并将该输入神经元及其连接关系数据写入上述映射单元601中；从上述存储单元602中读取处理后的权值，并将该权值写入上述第二输入缓存单元606，或者；

[0181] 从上述存储单元602中读取上述权值和该权值的连接关系数据，并将该权值及其连接关系数据写入上述映射单元601中；从上述存储单元602中读取处理后的输入神经元，并将该处理后的输入神经元写入上述第二输入缓存单元606；

[0182] 从上述输出缓存单元609中读取所述运算结果，并将该运算结果写入上述存储单元602中。

[0183] 其中，如图7所示，上述映射单元601包括：

[0184] 输入数据缓存单元6011，用于缓存第一输入数据，该第一输入数据包括至少一个输入神经元或者至少一个权值。

[0185] 连接关系缓存单元6012，用于缓存第一输入数据的连接关系数据，即上述输入神经元的连接关系数据或者上述权值的连接关系数据。

[0186] 其中，上述输入神经元的连接关系数据为用于表示该输入神经元中绝对值是否小于或者等于第一阈值的字符串，上述权值的连接关系数据为表示该权值绝对值是否小于或者等于上述第一阈值的字符串，或者为表示该权值对应的输入神经元和输出神经元之间是否有连接的字符串。该输入神经元的连接关系数据和权值的连接关系数据可以直接索引或者步长索引的形式表示。

[0187] 需要说明的是，上述直接索引和步长索引的描述可参见上述图1b所示实施例的相关描述。

[0188] 第四稀疏处理单元6013，用于根据所述第一输入数据的连接关系数据对所述第一输入数据进行处理，以得到处理后的第一输入数据，并将该处理后的第一输入数据存储到上述第一输入缓存单元中605。

[0189] 其中，当上述第一输入数据为至少一个输入神经元时，上述第四稀疏处理单元6013在一个时钟周期处理一个输入神经元和一个连接关系，即在一个时钟周期从S1个输入神经元中选择一个有效的输入神经元，S1为大于1的整数。

[0190] 在一种可行的实施例中，上述第四稀疏处理单元6013在一个时钟周期处理多个输入神经元和多个连接关系数据，即一个时钟周期从S1个输入神经元中选出有效的S2个输入数据，上述S2为大于0且小于或者等于该S1的整数。

[0191] 举例说明，如图8所示，上述输入神经元为i1，i2，i3和i4，以直接索引的形式表示的连接关系数据为“1011”，并且上述第四稀疏处理单元6013在一个时钟周期可从4个输入神经元选择1个有连接(即有效)的输入神经元。上述第四稀疏处理单元6013从上述输入数据缓存单元6011和上述连接关系缓存单元6012中分别获取上述输入神经元i1，i2，i3和i4和上述连接关系数据“1011”后，上述第四稀疏处理单元6013根据该连接关系数据“1011”从上述输入神经元i1，i2，i3和i4选取有连接的输入神经元i1，i3和i4。由于上述第四稀疏处理单元6013在一个时钟周期可从4个输入神经元选择1个有连接(即有效)的输入神经元，该第四稀疏处理单元6013在三个时钟周期内依次输出输入神经元i1，i3和i4，如图8所示。上述第四稀疏处理单元6013将上述输入神经元i1，i3和i4存储到上述第一输入缓存单元605

中。

[0192] 再举例说明，如图9所示，输入神经元为i1，i2，i3和i4，以直接索引的形式表示的连接关系数据有两组，分别为“1011”和“0101”，上述第四稀疏处理单元6013在一个时钟周期可从4个输入神经元中选择2个有连接(即有效)的输入神经元。上述第四稀疏处理单元6013根据上述连接关系数据“1011”从上述输入神经元i1，i2，i3和i4中选择有连接的输入神经元i1，i3和i4；根据上述连接关系数据“0101”从上述输入神经元i1，i2，i3和i4中选择有连接的输入神经元i2和i4。由于上述第四稀疏处理单元6013在一个时钟周期可从4个输入神经元选择2个有连接(即有效)的输入神经元，对于连接关系数据“1011”，该第四稀疏处理单元6013在第一个时钟周期从选择输入神经元i1和i3，并将该输入神经元i1和i3存储到上述第一输入缓存单元606中，在第二个时钟周期从选择输入神经元i4，并将该输入神经元i4存储到上述第一输入缓存单元606中；对于连接关系数据“0101”，该第四稀疏处理单元6013在一个时钟周期从选择输入神经元i2和i4，如图9所示。上述第四稀疏处理单元6013将上述输出神经元i2和i4和存储到上述第一输入缓存单元605中。

[0193] 举例说明，如图10所示，输入数据为输入神经元i1，i2，i3和i4，以步长索引的形式表示的连接关系数据为“021”，并且上述第四稀疏处理单元6013在一个时钟周期可从4个输入神经元选择1个有连接(即有效)的输入神经元。上述第四稀疏处理单元6013从上述输入数据缓存单元6011和上述连接关系缓存单元6012中分别获取上述输入神经元i1，i2，i3和i4和上述连接关系数据“021”后，上述第四稀疏处理单元6013根据该连接关系数据“1011”从上述输入神经元i1，i2，i3和i4选取有连接的输入神经元i1，i3和i4。由于上述第四稀疏处理单元6013在一个时钟周期可从4个输入神经元选择1个有连接(即有效)的输入神经元，该第四稀疏处理单元6013在三个时钟周期内依次输出输入神经元i1，i3和i4，如图10所示。上述第四稀疏处理单元6013将上述输入神经元i1，i3和i4存储到上述第一输入缓存单元605中。

[0194] 再举例说明，如图11所示，输入数据为输入神经元i1，i2，i3和i4，以步长索引的形式表示的连接关系数据有两组，分别为“021”和“22”，上述第四稀疏处理单元6013在一个时钟周期可从4个输入神经元中选择2个有连接(即有效)的输入神经元。上述第四稀疏处理单元6013根据上述连接关系数据“021”从上述输入神经元i1，i2，i3和i4中选择有连接的输入神经元i1，i3和i4；根据上述连接关系数据“22”从上述输入神经元i1，i2，i3和i4中选择有连接的输入神经元i2和i4。由于上述第四稀疏处理单元6013在一个时钟周期可从4个输入神经元选择2个有连接(即有效)的输入神经元，对于连接关系数据“021”，该第四稀疏处理单元6013在第一个时钟周期从选择输入神经元i1和i3，并将该输入神经元i1和i3存储到上述第一输入缓存单元606中。在第二个时钟周期从选择输入神经元i4并将该输入神经元i4存储到上述第一输入缓存单元606中；对于连接关系数据“22”，该第四稀疏处理单元6013在一个时钟周期从选择输入神经元i2和i4并输出，如图11所示。上述第四稀疏处理单元6013将上述输入神经元i2和i4存储到上述第一输入缓存单元605中。

[0195] 在一种可行的实施例中，上述输入数据缓存单元6011用于缓存的第一输入数据包括至少一个权值，上述连接关系缓存单元6012缓存的数据为上述权值的连接关系数据，且上述至少一个权值的绝对值均大于第一阈值时，上述第四稀疏处理单元6013根据上述权值的连接关系数据，将没有连接关系的输入神经元和输出神经元之间的权值的值置为0，并将

该值为0的权值和上述至少一个权值存储到上述第二输入缓存单元606中。

[0196] 举例说明，权值的形式为w_{ij}，表示第i个输入神经元与第j个输出神经元之间的权值。假设输入神经元包括i1，i2，i3和i4，输出神经元包括o1，上述第一输入数据(权值)为w_{11}，w_{31}，w_{41}，上述第一输入数据的连接关系数据(即上述权值的连接关系数据)以直接索引的形式表示，为1011，上述第四稀疏处理单元6013根据上述第二输入数据确定上述输入神经元i2与上述输出神经元o1之间没有连接，上述第四稀疏处理单元6013将该上述输入神经元i2与上述输出神经元o1之间的权值w_{21}的值置为0，并将w_{11}，w_{21}(0)，w_{31}，w_{41}存储到上述第二输入缓存单元606中。

[0197] 上述第一输入缓存单元605，用于缓存上述处理后的输入神经元。

[0198] 上述第二输入缓存单元606，用于缓存从上述存储单元602中读取的处理的权值。

[0199] 在一种可行的实施例中，当上述第一输入数据为至少一个权值时，上述第四稀疏处理单元6013在一个时钟周期处理一个权值和一个连接关系，即在一个时钟周期从S3个权值中选择一个有效的权值，该S3为大于1的整数。

[0200] 可选地，上述第四稀疏处理单元6013在一个时钟周期处理多个权值和多个连接关系数据，即一个时钟周期从S3个权值中选出有效的S4个权值，上述S4为大于0且小于或者等于该S3的整数。

[0201] 上述第一输入缓存单元605，用于缓存上述处理后的权值。

[0202] 上述第二输入缓存单元606，用于缓存从上述存储单元602中读取的处理的输入神经元。

[0203] 需要说明的是，上述相关描述可参见图8-图11的相关描述，在此不再叙述。

[0204] 可选地，所述映射单元601对所述第一输入数据进行处理之前，所述映射单元601还用于：

[0205] 对所述至少一个输入神经元进行分组，以得到M组输入神经元，所述M为大于或者等于1的整数；

[0206] 判断所述M组输入神经元的每一组输入神经元是否满足第一预设条件，所述第一预设条件包括一组输入神经元中绝对值小于或者等于第三阈值的输入神经元的个数小于或者等于第四阈值；

[0207] 当所述M组输入神经元任意一组输入神经元不满足所述第一预设条件时，将该组输入神经元删除；

[0208] 对所述至少一个权值进行分组，以得到N组权值，所述N为大于或者等于1的整数；

[0209] 判断所述N组权值的每一组权值是否满足第二预设条件，所述第二预设条件包括一组权值中绝对值小于或者等于第五阈值的权值的个数小于或者等于第六阈值；

[0210] 当所述N组权值任意一组权值不满足所述第二预设条件时，将该组权值删除。

[0211] 需要说明的是，上述相关描述可参见图1所示实施例中的相关描述，在此不再叙述。

[0212] 需要说明的是，第一阈值、第二阈值、第三阈值、第四阈值、第五阈值和第六阈值可均存储在上述存储单元602或者第一输出缓存单元605中；上述第一阈值、第二阈值、第三阈值、第四阈值、第五阈值和第六阈值中部分阈值存储在上述存储单元602、部分阈值存储在上述第一输出缓存单元605中。

[0213] 上述指令控制单元607，用于从上述指令缓存单元604中获取神经网络指令后，将该神经网路指令译码成运算单元608执行的微指令。

[0214] 上述运算单元608，用于从上述第一输入缓存605和上述第二输入缓存606中获取上述处理后的输入神经元和处理后的权值后，根据从上述微指令对上述处理后的权值和处理后的输入神经元进行人工神经网络运算，并将运算结果存储到上述输出缓存单元609中。

[0215] 上述输出缓存单元609，用于缓存上述运算单元608进行人工神经网络运算得到的运算结果。

[0216] 需要指出的是，上述指令缓存单元604、上述第一输入缓存单元605、上述第二输入缓存单元606和上述输出缓存单元609均可为片上缓存。

[0217] 需要说明的是，上述图1b、图5和图6所示实施例中的片上缓存是位于神经网络运算模块和内存之间的临时存储器，它的容量比内存小，但是交换速度快。片上缓存中的数据是内存中数据的一小部分，这一小部分数据是神经网络运算模块即将要访问的数据，当神经网络运算模块需要读写数据时，就可以直接访问片上缓存，从而加快读写数据的速度。

[0218] 需要说明的是，上述图1b、图5和图6所示的实施例中的权值的连接关系数据的表示方式除了直接索引和步长索引之外，还可为以下几种情况：

[0219] 方式一：列表的列表(List of Lists，LIL)

[0220] 以LIL的形式表示具体是将上述权值矩阵的每一行的非零权值的信息存储在一个列表中，该列表中的每个记录包括非零权值的列索引及该非零权值的值。

[0221] 举例说明，假设上述权值矩阵为$\begin{bmatrix} x1 & 0 \\ 0 & x4 \\ x2 & x5 \\ x3 & 0 \end{bmatrix}$，则该权值矩阵的连接关系数据用LIL的形式表示为((1,x1),(2,x4),((1,x2),(2,x5)),(1,x3))。该连接关系数据中有4个列表，表示该权值矩阵对应的输入神经元的数量为4个，分别为i1，i2，i3和i4。上述LIL中列表中最多有两个记录，由此可知该权值矩阵对应的输出神经元个数为2，分别为o1和o2。上述第一个列表中的记录(1,x1)表示输入神经元i1与输出神经元o1之间的权值为x1，上述第二个列表中的记录(2,x4)表示输入神经元i2与输出神经元o2之间的权值为x4，上述第三个列表中的记录(1,x2)表示输入神经元i3与输出神经元o1之间的权值为x2，记录(2,x5)表示输入神经元i3与输出神经元o1之间的权值为x5，上述第四个列表中的记录(1,x3)表示输入神经元i4与输出神经元o1之间的权值为x3。因此由上述LIL可得到如图12所示的神经网络结构。

[0222] 对于上述输出神经元o1，上述映射单元输出权值x1，x2和x3，该权值x1，x2和x3分别对应输入神经元i1，i3和i4；对于上述输出神经元o2，上述映射单元输出权值x4和x5，该权值x4和x5分别对应输入神经元i2和i3。上述映射单元将上述权值x1，x2和x3与x4和x5存储到上述第一输入缓存单元中。

[0223] 上述以LIL的形式表示上述权值的连接关系数据的优点在于简单，可快速构造矩阵，方便修改(按照列索引的大小顺序存储记录时)，支持灵活的切片操作。

[0224] 方式二：坐标列表(Coordinate list，COO)

[0225] 该坐标列表为由至少一个元组组成的列表，该元组包括非零权值在上述权值矩阵中的行号，列号和该非零权值的值组成的。每个元组表示序号为行号的输入神经元与序号

为列号的输出神经元之间的权值为该元组对应的非零权值。并且坐标列表的元组中的最大行号值为权值矩阵对应的输入神经元的个数，最大列号值为权值矩阵对应的输出神经元的个数。

[0226] 换句话说，上述坐标列表中每个元组表示非零权值在权值矩阵的位置信息。

[0227] 举例说明，假设上述权值矩阵为$\begin{bmatrix} x1 & 0 \\ 0 & x4 \\ x2 & x5 \\ x3 & 0 \end{bmatrix}$，则该权值矩阵的连接关系数据以COO的形式表示为(1,1,x1)，(2,2,x4)，(3,1,x2)，(3,2,x5)，(4,1,x3)，该COO的元组中最大行号值为4和最大的列号值为2，该权值矩阵对应的输入神经元个数为4和输出神经元的个数为2，分别为输入神经元i1，i2，i3、i4和输出神经元o1，o2，由元组(1,1,x1)可知输入神经元i1与输出神经元o1之间的权值为x1，由元组(2,2,x4)可知输入神经元i2与输出神经元o2之间的权值为x4，由元组(3,1,x2)可知输入神经元i3与输出神经元o1之间的权值为x2，由元组(3,2,x5)可知输入神经元i3与输出神经元o2之间的权值为x5，由元组(4,1,x3)可知输入神经元i4与输出神经元o1之间的权值为x3。由上述坐标列表可得到如图12所示的神经网络结构。

[0228] 对于上述输出神经元o1，上述映射单元输出权值x1，x2和x3，该权值x1，x2和x3分别对应输入神经元i1，i3和i4；对于上述输出神经元o2，上述映射单元输出权值x4和x5，该权值x4和x5分别对应输入神经元i2和i3。上述映射单元将上述权值x1，x2和x3、x4和x5存储到上述第一输入缓存单元中。

[0229] 上述以COO的形式表示上述权值的连接关系数据的优点在于简单，可以快速构建矩阵，方便修改。这种方法在矩阵特别稀疏的时候最适用，不管一个矩阵有多么巨大，若它只有一个非零元素，使用COO只需要3个数字，配合原矩阵的大小即可重建原矩阵，支持快速地与其他格式互相转化。

[0230] 方式三：压缩稀疏行(Compressed Sparse Row，CSR)

[0231] 采用CSR的形式是把上述权值矩阵行的信息压缩存储了，只显式保留每行第一个非零权值的位置。将上述权值矩阵通过三个数组表示：

[0232] 上述第一数组存储上述权值矩阵中的所有非零权值的值，其顺序按照从左至右、从上到下的行遍历方式排列元素，该第一数组记作A。该第一数组的长度即权值矩阵中非零权值的个数；

[0233] 上述第二数组存储上述第一数组A中的每个元素分别在权值矩阵中的列索引(即列号)，因而第二数组的长度与数组A的长度相同，记此数组为JA。

[0234] 上述第三数组记作IA，该数组IA的长度为权值矩阵的行数加1。该数组IA中的元素累加存储上述权值矩阵中每一行非零权值的个数，具体可通过如下递归方法获取，并在该数组IA中的最后一个元素中保存整个权值矩阵中非零权值的个数

[0235] 若上述三个数组的序号与权值矩阵的序号从0开始，可以用如下的递归方法定义数组IA：

[0236] IA[0]＝0

[0237] IA[i]＝IA[i-1]+权值矩阵中第i-1行的非零权值个数(i>0)

[0238] 举例说明，假设上述权值矩阵为 $\begin{bmatrix} 0 & 0 & 0 & 0 \\ 5 & 8 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 6 & 0 & 0 \end{bmatrix}$，由于上述第一数组A存储权值矩阵的所有非零权值的值，其顺序按照从左到右，从上到下的行遍历方式排列元素，故该第一数组A＝[5,8,3,6]；第二数组JA存储上述数组A中每个元素分别在上述权值矩阵中的列索引(即列号)，则该数组JA＝[0,1,2,1]。在第三数组IA中累加存储上述权值矩阵中每一行的非零权值的个数，根据上述递归方式确定该数组IA＝[0,0,2,3,4]。

[0239] 由上述第一数组A可知上述权值矩阵包括4个非零权值，分别为5,3,8,6。由上述第二数组JA可知上述4个非零权值在上述权值矩阵中的列索引，即权值5在上述权值矩阵中的第一列，权值8在上述权值矩阵中的第二列，权值3在上述权值矩阵中的第三列，权值6在上述权值矩阵中的第二列，由上述第三数组IA及其定义可知上述权值矩阵的第一行没有非零权值，第二行有2个非零权值，第三行和第四行各有1个非零权值；由上述信息可得到上述权值矩阵以坐标列表的形式表示为：(1,0,5)，(1,1,8)，(2,2,3)，(3,1,6)，进一步可确定上述权值矩阵。由该权值矩阵的形式可知，该矩阵的第一行和第四列的元素的值均为0，因此可确定该矩阵对应的输入神经元为3个，分别为i2，i3和i4；该权值矩阵对应的输出神经元分别为o1，o2和o3。最终可确定上述输入神经元i2与输出神经元o1之间的权值为5，上述输入神经元i2与输出神经元o2之间的权值为8，上述输入神经元i3与输出神经元o3之间的权值为3，上述输入神经元i4与输出神经元o2之间的权值为6，最终可得到如图13所示的神经网络结构。

[0240] 对于上述输出神经元o1，上述映射单元输出权值5，其对应输入神经元i2；对于上述输出神经元o2，上述映射单元输出权值8和6，其分别对应输入神经元i2和i4；对于上述输出神经元o3，上述映射单元输出权值3，其对应输入神经元i3。上述映射单元将上述权值5、8、6和3存储到上述第一输入缓存单元中。

[0241] 上述以CSR的形式表示上述权值的连接关系数据与COO的形式表示相比压缩了行索引的信息，并且采用CSR形式在存储稀疏矩阵时非零元素平均使用的字节数最为稳定。

[0242] 方式四：压缩稀疏列(Compressed Sparse Column，CSC)

[0243] 采用CSC的形式是把上述权值矩阵列的信息压缩存储了，只显式保留每列第一个非零权值的位置。将上述权值矩阵用三个数组表示：

[0244] 上述第四数组存储上述权值矩阵中的所有非零权值的值，其顺序按照从上至下、从左到右的列遍历方式排列元素，该第四数组记作A’，其长度即权值矩阵中非零权值的个数；

[0245] 上述第五数组存储上述第一数组A’中的每个元素分别在权值矩阵中的行索引(即行号)，因而其长度与第一数组A’相同，记此数组为JA’。

[0246] 上述第六数组记作IA’，该数组的长度为上述权值矩阵的列数加1。该数组IA’中的元素累加存储上述权值矩阵中每一列非零权值的个数，具体可通过如下递归方法获取，并且在该数组IA’累加整个权值矩阵中每一列中非零权值的个数。

[0247] 若上述三个数组的序号与权值矩阵的序号从0开始，可以用如下的递归方法定义数组IA’：

[0248] IA’[0]＝0

[0249] IA’[j]＝IA’[j-1]+权值矩阵中第j-1列的非零权值个数(j>0)

[0250] 举例说明，假设上述权值矩阵为$\begin{bmatrix}4&0&0&2\\0&1&0&0\\0&0&5&7\\6&3&0&8\end{bmatrix}$，由于上述第四数组A’存储权值矩阵的所有非零权值的数，其顺序按照从左到右，从上到下的行遍历方式排列元素，故该数组A’＝[4,6,1,3,5,2,7,8]；上述第五数组JA’存储上述数组A’中每个元素分别在上述权值矩阵中的行索引(即行号)，则该数组JA’＝[0,3,1,3,2,0,2,3]；根据上述递归方式确定数组IA’＝[0,2,4,5,8]。

[0251] 由上述数组A’可知上述权值矩阵包括8个非零权值，分别为4,6,1,3,5,2,7,8。由上述数组JA’可知上述8个非零权值在上述权值矩阵中的行索引，即权值4在上述权值矩阵中的第一列，权值6在上述权值矩阵中的第4列，权值1在上述权值矩阵中的第二列，权值3在上述权值矩阵中的第四列，权值5在上述权值矩阵中的第三列，权值2在上述权值矩阵中的第一列，权值7在上述权值矩阵中的第三列，权值8在上述权值矩阵中的第四列，由上述数组IA’及其定义可知上述权值矩阵的第一列和第二列各有2个非零权值，第三列有1个非零权值，第四行有3个非零权值；由上述信息可得到上述权值矩阵以坐标列表的形式表示为：(0,0,4)，(3,0,6)，(1,1,1)，(3,1,3)，(2,2,5)，(0,3,2)，(2,3,7)，(3,3,8)，进一步可确定上述权值矩阵。由该权值矩阵的形式可知，该矩阵的每一行和每一列均有非零权值，因此可确定该矩阵对应的输入神经元为4个，分别为i1，i2，i3和i4；该权值矩阵对应的输出神经元分别为o1，o2，o3和o4。最终可确定上述输入神经元i1与输出神经元o1之间的权值为4，上述输入神经元i1与输出神经元o4之间的权值为2，上述输入神经元i2与输出神经元o2之间的权值为1，上述输入神经元i3与输出神经元o3之间的权值为5，上述输入神经元i3与输出神经元o4之间的权值为7，上述输入神经元i4与输出神经元o1之间的权值为6，上述输入神经元i4与输出神经元o2之间的权值为3，上述输入神经元i4与输出神经元o4之间的权值为8，最终可得到如图14所示的神经网络结构。

[0252] 对于上述输出神经元o1，上述映射单元输出权值5和4，其分别对应输入神经元i1和i4；对于上述输出神经元o2，上述映射单元输出权值1和3，其分别对应输入神经元i2和i4；对于上述输出神经元o3，上述映射单元输出权值5，其对应输入神经元i3；对于上述输出神经元o3，上述映射单元输出权值2,7和8，其对应输入神经元i1，i3和i4。上述映射单元将上述权值4,6,1,3,5,2,7和8存储到上述第一输入缓存单元中。

[0253] 上述以CSC的形式表示上述权值的连接关系数据与COO的形式表示相比压缩了列索引的信息，对于算术运算、列切片、矩阵与向量的点乘都很有效。

[0254] 方式五：(ELL Pack，ELL)

[0255] 该方式采用两个与权值矩阵的行数相同矩阵存储该权值矩阵中非零权值的信息。上述第一矩阵存储上述权值矩阵中非零权值的列号，上述第二矩阵存储上述权值矩阵中非零权值的值，行号就不存了，用自身所在的行来表示；这两个矩阵每一行都是从头开始放，如果没有元素了就用个结束标志(比如*)结束。

[0256] 举例说明，假设上述权值矩阵为$\begin{bmatrix} x1 & 0 \\ 0 & x4 \\ x2 & x5 \\ x3 & 0 \end{bmatrix}$，则该权值矩阵的连接关系数据用ELL的形式表示为：

[0257] 第一矩阵为：$\begin{bmatrix} 0 & * \\ 1 & * \\ 0 & 1 \\ 0 & * \end{bmatrix}$，第二矩阵为$\begin{bmatrix} x1 & 0 \\ x4 & 0 \\ x2 & x5 \\ x3 & 0 \end{bmatrix}$。

[0258] 由上述第一矩阵和第二矩阵的行数可知，上述权值矩阵对应的输入神经元的个数为4，分别为输入神经元i1，i2，i3和i4；由上述第一矩阵和第二矩阵的列数可知，上述权值矩阵对应的输出神经元的个数为2，分别为输出神经元o1和o2。根据上述第一矩阵和第二矩阵可知，上述输入神经元i1与输出神经元o1之间的权值为x1，输入神经元i2与输出神经元o2之间的权值为x4，输入神经元i3与输出神经元o1之间的权值为x2，输入神经元i3与输出神经元o2之间的权值为x5，输入神经元i4与输出神经元o1之间的权值为x3。由上述ELL表示的连接关系数据可得到如图12所示的神经网络结构。

[0259] 对于上述输出神经元o1，上述映射单元输出权值x1，x2和x3，该权值x1，x2和x3分别对应输入神经元i1，i3和i4；对于上述输出神经元o2，上述映射单元输出权值x4和x5，该权值x4和x5分别对应输入神经元i2和i3。上述映射单元将上述权值x1，x2和x3、x4和x5存储到上述第一输入缓存单元中。

[0260] 对于通过ELL方式表示的连接关系数据，当权值矩阵的某一行的非零元素多余其他行时，在第一矩阵中的结尾处会存在多个结束标志，浪费缓存资源。为了解决该问题，可采用方式六所示的方式表示上述连接关系数据。

[0261] 方式六：混合(Hybird，HYB)

[0262] 该方式可以看成上述ELL和COO方式的组合。采用COO的方式存储权值矩阵中某一行相对于其他行多出来的非零权值。采用ELL的方式存储权值矩阵中每一行最大相同数量的非零权值。

[0263] 假设上述权值矩阵为：$\begin{bmatrix} 1 & 7 & 0 & 0 \\ 0 & 2 & 8 & 0 \\ 5 & 0 & 3 & 9 \\ 0 & 6 & 0 & 4 \end{bmatrix}$，则上述ELL中的第三矩阵为$\begin{bmatrix} 0 & 1 \\ 1 & 2 \\ 0 & 2 \\ 1 & 3 \end{bmatrix}$，第四矩阵为$\begin{bmatrix} 1 & 7 \\ 2 & 8 \\ 5 & 3 \\ 6 & 4 \end{bmatrix}$；上述COO形式的元组为(2,3,9)。由上述第三矩阵和第四矩阵的行数可知，上述权值矩阵对应的输入神经元的个数为4，分别为

[0264] 输入神经元i1，i2，i3和i4；根据上述坐标列表中的列号(3)可知上述权值矩阵对应的输出神经元的个数为4，分别为输出神经元o1，o2，o3和o4。由上述第一矩阵，第二矩阵

和坐标列表可知:输入神经元i1与输出神经元o1之间的权值为1,输入神经元i1与输出神经元o2之间的权值为7,输入神经元i2与输出神经元o2之间的权值为2,输入神经元i2与输出神经元o3之间的权值为8,输入神经元i3与输出神经元o1之间的权值为5,输入神经元i3与输出神经元o3之间的权值为3,输入神经元i3与输出神经元o4之间的权值为9,输入神经元i4与输出神经元o2之间的权值为6,输入神经元i4与输出神经元o4之间的权值为4,可以得到如图15所示的神经网络结构。

[0265] 对于上述输出神经元o1,上述映射单元输出权值1和5,分别对应输入神经元i1和i3;对于上述输出神经元o2,上述映射单元输出权值7和2,分别对应输入神经元i1和i2;对于上述输出神经元o3,上述映射单元输出权值8和3,分别对应输入神经元i2和i3;对于上述输出神经元o4,上述映射单元输出权值9和4,分别对应输入神经元i3和i4。上述映射单元将上述权值1,5,7,2,8,3,9和4存储到上述第一输入缓存单元中。

[0266] 总而言之,这六种形式(LIL、COO、CSC、CSR、ELL、HYB)在稀疏度越高的情况下越能占用更少的存储空间。LIL根据具体实现结构的不同,占用稍多于2*nnz个存储单元,空间代价优于其他方法。如果非零元素数量小于行数\列数,那么使用COO比使用CSR/CSC更加经济,反之则使用CSR/CSC更加经济。如果每行的非零元素数目比较均匀,即矩阵中的每一行的非零元素个数差别不大,这样非零元素最多的行中的非零元素数目,与不均匀的矩阵中相应非零元素最多行相比,显然会更少,那么可以考虑使用ELL。在极端均匀的情况下,即每一行的非零元素个数都一样,ELL所占存储单元个数是2*nnz,比COO和CSR、CSC都要少。但是稀疏神经网络并不能保证有这样的特性。也许有某些特定的稀疏神经网络模型会有这样的特性,那么使用ELL比较好。对于矩阵中每一行稀疏元素个数较统一的情况,采用ELL形式的表示最佳,其次是HYB(ELL+COO)。

[0267] 在并行方面,COO是可以并行生成的,CSR与CSC的3个数组中的2个也是可以并行生成的。在做运算时,COO、LIL、ELL均可按行并行计算,而CSC、CSR、HYB则需要更多的预处理。

[0268] CSR擅长稀疏矩阵左乘向量,而CSC擅长于稀疏矩阵右乘向量转置。这两种表示形式可以通过转置互相转换。在神经网络的传播过程中可以使用这两种方法以及COO。ELL格式在进行稀疏矩阵-矢量乘积(sparse matrix-vector products)时效率最高。

[0269] 需要说明的是,上述非零权值还可以替换为大于第一预设阈值的权值。

[0270] 可选地,上述第一预设阈值可为0.5、1、1.2、1.5、2或者其他值。

[0271] 需要说明的是,上述图1b、图5和图6所示的实施例中的输入神经元的连接关系数据的表示方式除了直接索引和步长索引之外,还以以上述六种方式(LIL、COO、CSR、CSC、ELL、HYB)进行表示。

[0272] 当以上述六种方式表示上述输入神经元的连接关系数据时,上述非零权值可替换为非零输入神经元,上述权值矩阵可替换为输入神经元矩阵。

[0273] 进一步的,上述非零输入神经元可替换为大于第二预设阈值的输入神经元。

[0274] 可选地,上述第二预设阈值可为0.5、1、1.2、1.5、2或者其他值。上述第一预设阈值和上述第二预设阈值可以相同或者不同。

[0275] 需要说明的是,图1b、图5和图6所示的实施例中相关的连接关系数据(包括权值的连接关系数据和输入神经元的连接关系数据)可以采用高维动态数组,可以用链表等等表示。

[0276] 需要说明的是，上述神经网络运算模块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。上述神经网络运算模块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。

[0277] 需要指出的是，本发明实施例中提到的输入神经元和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输出神经元。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输出神经元，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。

[0278] 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。

[0279] 在一种可行的实施例中，本发明实施例提供了一种神经网络运算装置，该神经网络运算装置包括一个或多个如图1b、图5或者图6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；

[0280] 当所述神经网络运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；

[0281] 其中，多个所述神经网络运算模块通过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；多个所述神经网络运算模块共享同一控制系统或拥有各自的控制系统；多个所述神经网络运算模块共享内存或者拥有各自的内存；多个所述神经网络运算模块的互联方式是任意互联拓扑。

[0282] 该神经网络运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。

[0283] 在一种可行的实施例中，本发明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运算装置，通用互联接口和其他处理装置。

[0284] 上述神经网络运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16a，图16a为本发明实施例提供的一种组合处理装置的结构示意图。如图16a所示，该组合处理装置包括上述神经网络运算装置1601、通用互联接口1602和其他处理装置1603。

[0285] 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作为神经网络运算装置1601与外部数据和控制的接口，包括数据搬运，完成对本神经

网络运算装置的开启、停止等基本控制;其他处理装置1603也可以和神经网络运算装置1601协作共同完成运算任务。

[0286] 上述通用互联接口1602,用于在所述神经网络运算装置1601与其他处理装置1603间传输数据和控制指令。该神经网络运算装置1601从其他处理装置1603中获取所需的输入数据,写入神经网络运算装置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,写入神经网络运算装置1601片上的控制缓存;也可以读取神经网络运算装置1601的存储模块中的数据并传输给其他处理装置1603。

[0287] 可选的,如图16b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运算装置1601或其他处理装置1603的内部存储中无法全部保存的数据。

[0288] 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。

[0289] 在一种可行的实施例中,本发明实施例提供了一种神经网络芯片,该神经网络芯片包括如图1b、图5或者图6所示实施例所述的神经网络运算模块,或者上述神经网络运算装置或者上述组合处理装置。

[0290] 在一种可行的实施例中,本发明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。

[0291] 在一种可行的实施例中,本发明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。该板卡可用于众多通用或专用的计算系统环境或配置中。例如:个人计算机、服务器计算机、手持设备或便携式设备、平板型设备、智能家居、家电、多处理器系统、基于微处理器的系统、机器人、可编程的消费电子设备、网络个人计算机(personal computer,PC)、小型计算机、大型计算机、包括以上任何系统或设备的分布式计算环境等等。

[0292] 请参照图17,图17为本发明实施例提供的一种板卡的结构示意图。如图17所示,上述板卡17包括神经网络芯片封装结构171、第一电气及非电气连接装置172和第一基板(substrate)173。

[0293] 对于神经网络芯片封装结构171的具体结构不作限定,可选的,如图18所示,上述神经网络芯片封装结构171包括:神经网络芯片1711、第二电气及非电气连接装置1712、第二基板1713。

[0294] 本发明所涉及的神经网络芯片1711的具体形式不作限定,上述的神经网络芯片1711包含但不限于将神经网络处理器集成的神经网络晶片上,上述晶片可以由硅材料、锗材料、量子材料或分子材料等制成。根据实际情况(例如:较严苛的环境)和不同的应用需求可将上述神经网络晶片进行封装,以使神经网络晶片的大部分被包裹住,而将神经网络晶片上的引脚通过金线等导体连到封装结构的外边,用于和更外层进行电路连接。

[0295] 本发明对于第一基板173和第二基板1713的类型不做限定,可以是印制电路板(printed circuit board,PCB)或(printed wiring board,PWB),还可能为其它电路板。对PCB的制作材料也不做限定。

[0296] 本发明所涉及的第二基板1713用于承载上述神经网络芯片1711,通过第二电气及非电气连接装置1712将上述的神经网络芯片1711和第二基板1713进行连接得到的神经网

络芯片封装结构171,用于保护神经网络芯片1711,便于将神经网络芯片封装结构171与第一基板173进行进一步封装。

[0297] 对于上述具体的第二电气及非电气连接装置1712的封装方式和封装方式对应的结构不作限定,可根据实际情况和不同的应用需求选择合适的封装方式并进行简单地改进,例如:倒装芯片球栅阵列封装(Flip Chip Ball Grid Array Package,FCBGAP),薄型四方扁平式封装(Low-profile Quad Flat Package,LQFP)、带散热器的四方扁平封装(Quad Flat Package with Heat sink,HQFP)、无引脚四方扁平封装(Quad Flat Non-lead Package,QFN)或小间距四方扁平式封装(Fine-pitch Ball Grid Package,FBGA)等封装方式。

[0298] 倒装芯片(Flip Chip),适用于对封装后的面积要求高或对导线的电感、信号的传输时间敏感的情况下。除此之外可以用引线键合(Wire Bonding)的封装方式,减少成本,提高封装结构的灵活性。

[0299] 球栅阵列(Ball Grid Array),能够提供更多引脚,且引脚的平均导线长度短,具备高速传递信号的作用,其中,封装可以用引脚网格阵列封装(Pin Grid Array,PGA)、零插拔力(Zero Insertion Force,ZIF)、单边接触连接(Single Edge Contact Connection,SECC)、触点阵列(Land Grid Array,LGA)等来代替。

[0300] 可选的,采用倒装芯片球栅阵列(Flip Chip Ball Grid Array)的封装方式对神经网络芯片1711和第二基板1713进行封装,具体的神经网络芯片封装结构171的示意图可参照图19。如图19所示,上述神经网络芯片封装结构包括:神经网络芯片21、焊盘22、焊球23、第二基板24、第二基板24上的连接点25、引脚26。

[0301] 其中,焊盘22与神经网络芯片21相连,通过在焊盘22和第二基板24上的连接点25之间焊接形成焊球23,将神经网络芯片21和第二基板24连接,即实现了神经网络芯片21的封装。

[0302] 引脚26用于与封装结构的外部电路(例如,神经网络处理器板卡17上的第一基板173)相连,可实现外部数据和内部数据的传输,便于神经网络芯片21或神经网络芯片21对应的神经网络处理器对数据进行处理。对于引脚的类型和数量本发明也不作限定,根据不同的封装技术可选用不同的引脚形式,并遵从一定规则进行排列。

[0303] 可选的,上述神经网络芯片封装结构还包括绝缘填充物,置于焊盘22、焊球23和连接点25之间的空隙中,用于防止焊球与焊球之间产生干扰。

[0304] 其中,绝缘填充物的材料可以是氮化硅、氧化硅或氧氮化硅;干扰包含电磁干扰、电感干扰等。

[0305] 可选的,上述神经网络芯片封装结构还包括散热装置,用于散发神经网络芯片21运行时的热量。其中,散热装置可以是一块导热性良好的金属片、散热片或散热器,例如,风扇。

[0306] 举例来说,如图20所示,上述神经网络芯片封装结构171包括:神经网络芯片21、焊盘22、焊球23、第二基板24、第二基板24上的连接点25、引脚26、绝缘填充物27、散热膏28和金属外壳散热片29。其中,散热膏28和金属外壳散热片29用于散发神经网络芯片21运行时的热量。

[0307] 可选的,上述神经网络芯片封装结构171还包括补强结构,与焊盘22连接,且内埋

于焊球23中,以增强焊球23与焊盘22之间的连接强度。

[0308] 其中,补强结构可以是金属线结构或柱状结构,在此不做限定。

[0309] 本发明对于第一电气及非电气装置172的具体形式也不作限定,可参照第二电气及非电气装置1712的描述,即通过焊接的方式将神经网络芯片封装结构171进行封装,也可以采用连接线连接或插拔方式连接第二基板1713和第一基板173的方式,便于后续更换第一基板173或神经网络芯片封装结构171。

[0310] 可选的,第一基板173包括用于扩展存储容量的内存单元的接口等,例如:同步动态随机存储器(Synchronous Dynamic Random Access Memory,SDRAM)、双倍速率同步动态随机存储器(Double Date Rate SDRAM,DDR)等,通过扩展内存提高了神经网络处理器的处理能力。

[0311] 第一基板173上还可包括快速外部设备互连总线(Peripheral Component Interconnect-Express,PCI-E或PCIe)接口、小封装可热插拔(Small Form-factor Pluggable,SFP)接口、以太网接口、控制器局域网总线(Controller Area Network,CAN)接口等等,用于封装结构和外部电路之间的数据传输,可提高运算速度和操作的便利性。

[0312] 将神经网络处理器封装为神经网络芯片1711,将神经网络芯片1711封装为神经网络芯片封装结构171,将神经网络芯片封装结构171封装为板卡17,可填补目前神经网络的空缺,通过板卡上的接口(插槽或插芯)与外部电路(例如:计算机主板)进行数据交互,即直接通过使用板卡17实现神经网络处理器的功能,并保护神经网络芯片1711。且板卡17上还可添加其他模块,提高了神经网络处理器的应用范围和运算效率。

[0313] 在一种可行的实施例中,本发明实施例提供了一种电子装置,该电子装置包括上述板卡。

[0314] 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。

[0315] 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。

[0316] 参见图21,图21为本发明实施例提供的一种神经网络运算方法的流程示意图。如图21所示,该方法包括:

[0317] S2101、神经网络运算模块对输入数据进行处理,以得到处理后的输入数据。

[0318] 其中,所述输入数据包括至少一个输入神经元和/或至少一个权值,所述对输入数据进行处理之前,所述方法还包括:

[0319] 对所述至少一个输入神经元进行分组,以得到M组输入神经元,所述M为大于或者等于1的整数;

[0320] 判断所述M组输入神经元的每一组输入神经元是否满足第一预设条件,所述第一预设条件包括一组输入神经元中绝对值小于或者等于第三阈值的输入神经元的个数小于或者等于第四阈值;

[0321] 当所述M组输入神经元任意一组输入神经元不满足所述第一预设条件时,将该组输入神经元删除;

[0322] 对所述至少一个权值进行分组，以得到N组权值，所述N为大于或者等于1的整数；

[0323] 判断所述N组权值的每一组权值是否满足第二预设条件，所述第二预设条件包括一组权值中绝对值小于或者等于第五阈值的权值的个数小于或者等于第六阈值；

[0324] 当所述N组权值任意一组权值不满足所述第二预设条件时，将该组权值删除。

[0325] 可选地，所述输入数据包括第一输入数据和第二输入数据，所述处理后的输入数据包括处理后的第一输入数据和处理后的第二输入数据，所述对输入数据进行处理，以得到处理后的输入数据，包括：

[0326] 对所述第二输入数据进行处理，以得到第一连接关系数据和处理后的第二输出数据；

[0327] 根据所述第一连接关系数据对所述第一输入数据进行处理，以得到处理后的第二输入数据，

[0328] 其中，当所述第一输入数据为输入神经元，所述第二输入数据为权值时，所述第一连接关系数据为所述权值的连接关系数据；当所述第一输入数据为权值，所述第二输入数据为输入神经元时，所述第一连接关系数据为输入神经元的连接关系数据。

[0329] 可选地，所述输入数据包括输入神经元和权值，所述处理后的输入数据包括处理后的输入神经元和处理后的权值，所述对输入数据进行处理，以得到处理后的输入数据，包括：

[0330] 根据所述输入神经元和所述权值获取所述输入神经元的连接关系数据和所述权值的连接关系数据；

[0331] 对所述输入神经元的连接关系数据和所述权值的连接关系数据进行处理，以得到第二连接关系数据，

[0332] 根据所述第二连接关系数据对所述输入神经元和所述权值进行处理，以得到所述处理后的输入神经元和所述处理后的权值。

[0333] 可选地，所述输入神经元的连接关系数据和所述权值的连接关系数据以直接索引的形式表示，所述对所述输入神经元的连接关系数据和所述权值的连接关系数据进行处理，以得到第二连接关系数据，包括：

[0334] 对所述输入神经元的连接关系数据和所述权值的连接关系数据进行与操作，以得到所述第三连接关系数据。

[0335] 可选地，所述对所述输入神经元的连接关系数据和所述权值的连接关系数据进行处理，以得到第二连接关系数据，包括：

[0336] 当所述输入神经元的连接关系数据以直接索引的形式表示，所述权值的连接关系数据以步长索引的形式表示时，将所述权值的连接关系数据转换成以直接索引的形式表示的连接关系数据；

[0337] 当所述权值的连接关系数据以直接索引的形式表示，所述输入神经元的连接关系数据以步长索引的形式表示时，将所述输入神经元的连接关系数据转换成以直接索引的形式表示的连接关系数据；

[0338] 对所述输入神经元的连接关系数据和所述权值的连接关系数据进行与操作，以得到所述第三连接关系数据。

[0339] 可选地，当所述输入神经元的连接关系数据和所述权值的连接关系数均以步长的

形式表，且表示所述权值的连接关系数据和所述输入神经元的连接关系数据的字符串是按照物理地址由低到高的顺序存储时，所述对所述输入神经元的连接关系数据和所述权值的连接关系数据进行处理，以得到第二连接关系数据，包括：

[0340] 将所述输入神经元的连接关系数据的字符串中的每一个元素与存储物理地址低于该元素存储的物理地址的元素进行累加，得到的新的元素组成第三连接关系数据；同理，对所述权值的连接关系数据的字符串进行同样的处理，得到第四连接关系数据；

[0341] 从所述第三连接关系数据的字符串和所述第四连接关系数据的字符串中，选取相同的元素，按照元素值从小到大的顺序排序，组成新的字符串；

[0342] 将所述新的字符串中每一个元素与其相邻且值小于该元素值的元素进行相减，得到的元素组成所述第三连接关系数据。

[0343] 可选地，当表示所述权值的连接关系数据和所述输入神经元的连接关系数据的字符串是按照物理地址由低到高的顺序存储时所述对所述输入神经元的连接关系数据和所述权值的连接关系数据进行处理，以得到第二连接关系数据，包括：

[0344] 当所述输入神经元的关系数据是以步长索引的形式表示，所述权值的连接关系数据是以直接索引的形式表示时，将所述权值的连接关系数据转换成以步长索引的形式表示的连接关系数据；

[0345] 当所述权值的关系数据是以步长索引的形式表示，所述输入神经元的连接关系数据是以直接索引的形式表示时，将所述输入神经元的连接关系数据转换成以步长索引的形式表示的连接关系数据；

[0346] 将所述第一连接关系数据的字符串中的每一个元素与存储物理地址低于该元素存储的物理地址的元素进行累加，得到的新的元素组成第四连接关系数据；同理，对所述第二连接关系数据的字符串进行同样的处理，得到第五连接关系数据；

[0347] 从所述第四连接关系数据的字符串和所述第五连接关系数据的字符串中，选取相同的元素，按照元素值从小到大的顺序排序，组成新的字符串；

[0348] 将所述新的字符串中每一个元素与其相邻且值小于该元素值的元素进行相减，得到的元素组成所述第三连接关系数据。

[0349] 可选地，所述对输入数据进行处理，以得到处理后的输入数据，包括：

[0350] 当所述输入数据包括输入神经元和所述输入神经元的连接关系数据时，根据所述输入神经元的连接关系数据对所述输入神经元进行处理，以得到处理后的输入神经元；

[0351] 当所述输入数据包括权值和所述权值的连接关系数据时，根据所述权值的连接关系数据对所述权值进行处理，以得到处理后的权值。

[0352] 其中，所述输入神经元的连接关系数据和所述权值的连接关系数据以直接索引或者步长索引的形式表示；

[0353] 当所述输入神经元的连接关系数据以直接索引的形式表示时，该连接关系数据为由0和1组成的字符串，0表示所述输入神经元的值的绝对值小于或者等于第一阈值，1表示所述输入神经元的值的绝对值大于所述第一阈值；

[0354] 当所述输入神经元的连接关系数据以步长索引形式表示时，该连接关系数据为绝对值大于所述第一阈值的输入神经元与上一个绝对值大于所述第一阈值的输入神经元之间的距离值组成的字符串；

[0355] 当所述权值的连接关系数据以直接索引的形式表示时，该连接关系数据为由0和1组成的字符串，0表示该权值的绝对值小于或者等于第二阈值，即该权值对应的输入神经元与输出神经元之间没有连接，1表示该权值的绝对值大于上述第二阈值，即该权值对应的输入神经元与输出神经元之间有连接；以直接索引形式表示权值的连接关系数据有两种表示顺序：以每个输出神经元与所有输入神经元的连接状态组成一个0和1的字符串来表示所述权值的连接关系数据；或者每个输入神经元与所有输出神经元的连接状态组成一个0和1的字符串来表示所述权值的连接关系数据；

[0356] 当所述权值的连接关系数据以步长索引的形式表示时，该连接关系数据为与输出神经元有连接的输入神经元的与上一个与该输出神经元有连接的输入神经元之间的距离值组成的字符串。

[0357] S2102、神经网络运算模块获取神经运算指令，将所述神经运算指令译码成微指令。

[0358] S2103、神经网络运算模块根据所述微指令对所述处理后的输入数据进行人工神经网络运算，以得到运算结果。

[0359] 需要说明的是，上述步骤S2101-S2103的描述可参见上述图1b、图5和图6所示实施例的相关描述，在此不再叙述。

[0360] 本发明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。

[0361] 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，本发明并不受所描述的动作顺序的限制，因为依据本发明，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优选实施例，所涉及的动作和模块并不一定是本发明所必须的。

[0362] 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。

[0363] 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。

[0364] 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。

[0365] 另外，在本发明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，也可以采用软件功能单元的形式实现。

[0366] 所述集成的单元如果以软件功能单元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,本发明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)执行本发明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。

[0367] 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。

[0368] 以上对本发明实施例进行了详细介绍,本文中应用了具体个例对本发明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发明的方法及其核心思想;同时,对于本领域的一般技术人员,依据本发明的思想,在具体实施方式及应用范围上均会有改变之处,综上上述,本说明书内容不应理解为对本发明的限制。

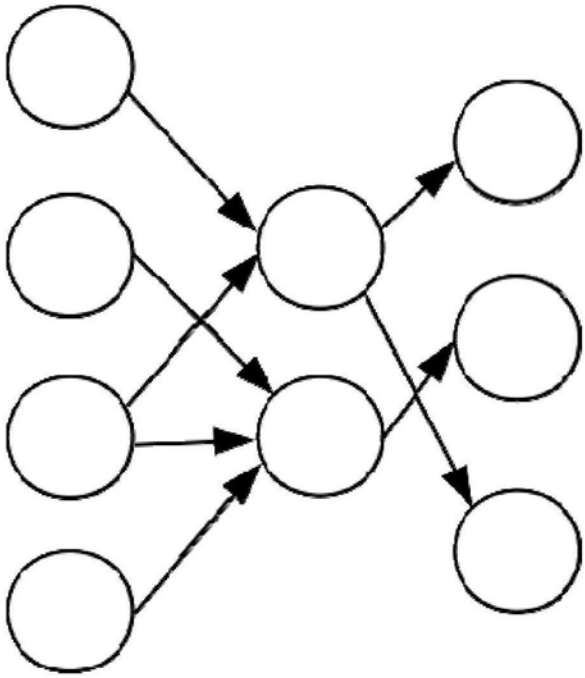


图1a

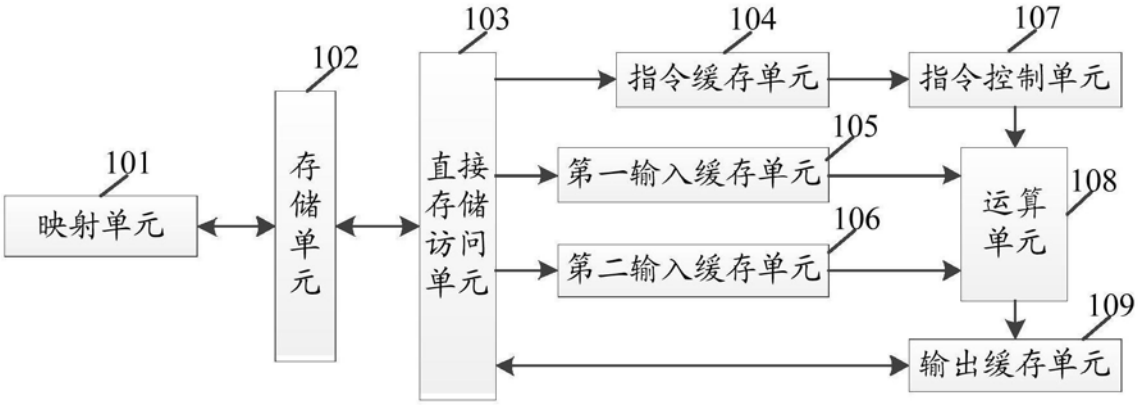

101
映射单元
102
存储单元
103
直接存储访问单元
104
指令缓存单元
105
第一输入缓存单元
106
第二输入缓存单元
107
指令控制单元
108
运算单元
109
输出缓存单元

图1b

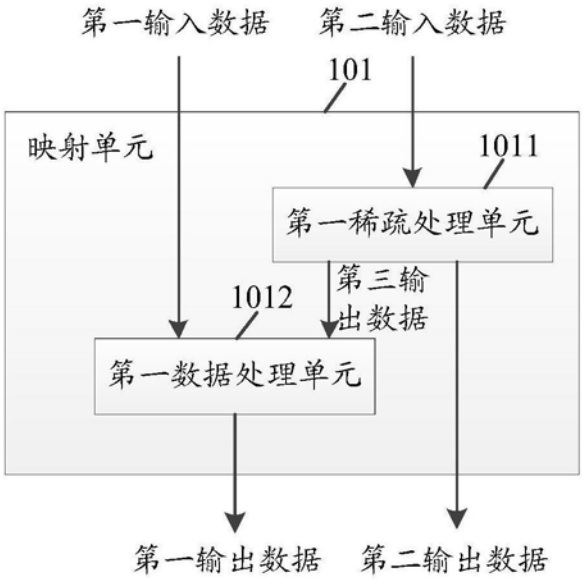

第一输入数据
第二输入数据
101
映射单元
1011
第一稀疏处理单元
第三输出数据
1012
第一数据处理单元
第一输出数据
第二输出数据

图2

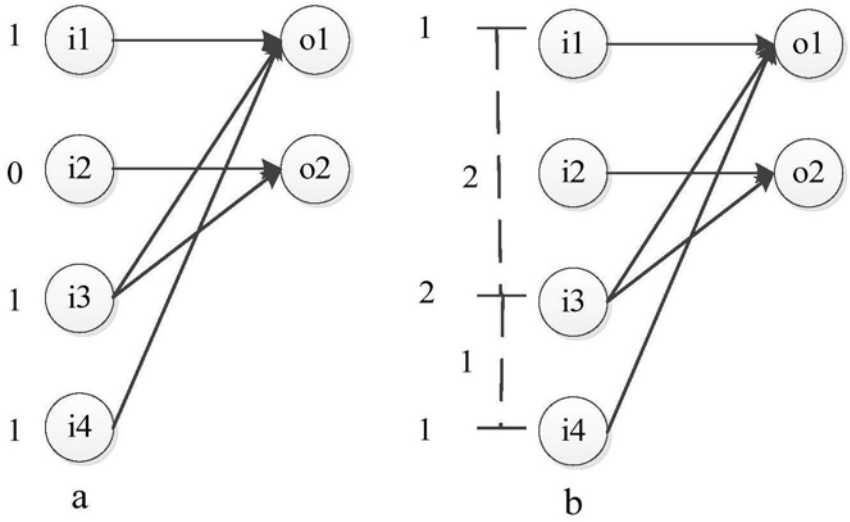

1
i1
o1
0
i2
o2
1
i3
1
i4
a
1
i1
o1
2
i2
o2
2
i3
1
1
i4
b

图3

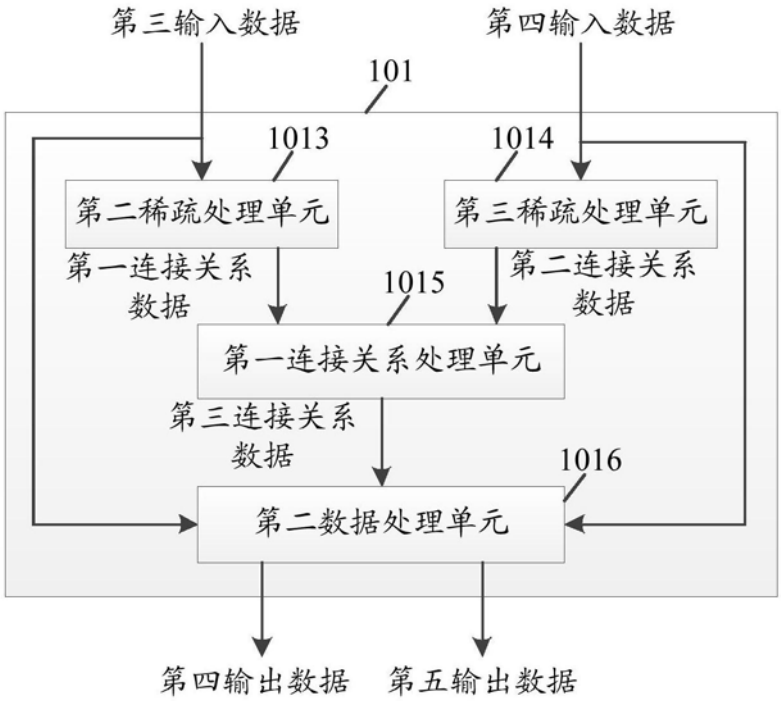

第三输入数据
第四输入数据
101
1013
1014
第二稀疏处理单元
第三稀疏处理单元
第一连接关系数据
第二连接关系数据
1015
第一连接关系处理单元
第三连接关系数据
1016
第二数据处理单元
第四输出数据
第五输出数据

图4

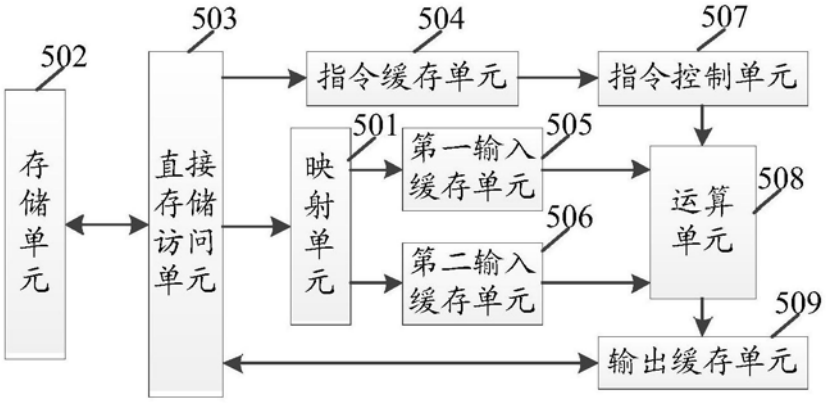

503
504
507
502
指令缓存单元
指令控制单元
501
505
存储单元
直接存储访问单元
映射单元
第一输入缓存单元
运算单元
508
506
第二输入缓存单元
509
输出缓存单元

图5

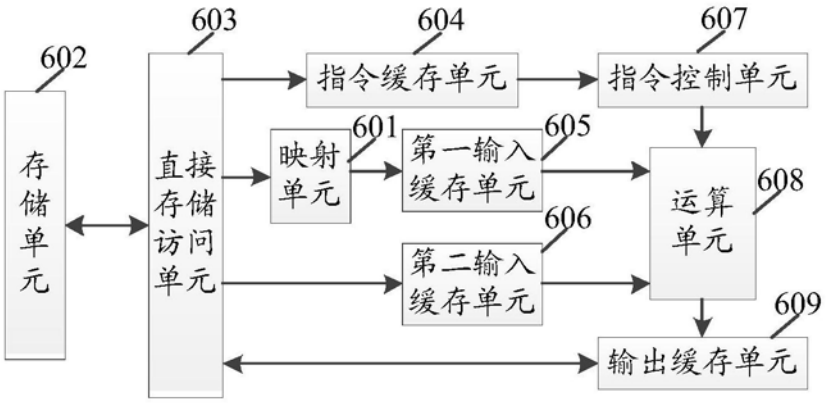

603
604
607
602
指令缓存单元
指令控制单元
存储单元
直接存储访问单元
601
映射单元
第一输入缓存单元
605
运算单元
608
606
第二输入缓存单元
609
输出缓存单元

图6

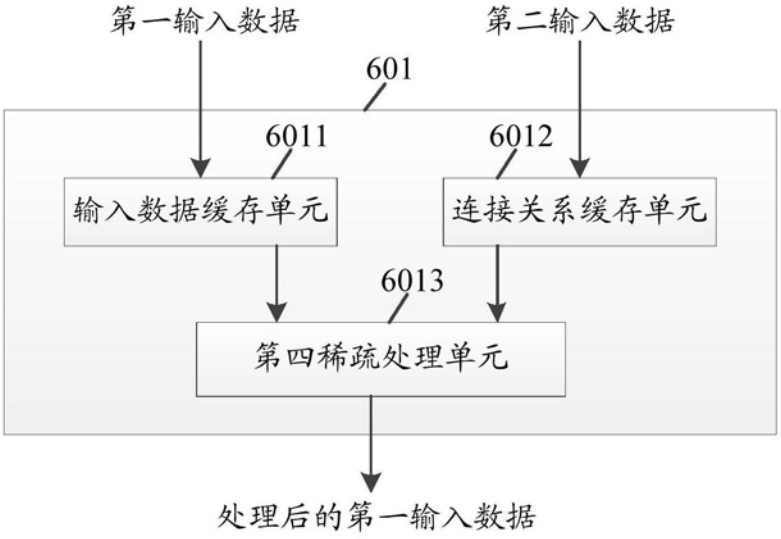

第一输入数据
第二输入数据
601
6011
6012
输入数据缓存单元
连接关系缓存单元
6013
第四稀疏处理单元
处理后的第一输入数据

图7

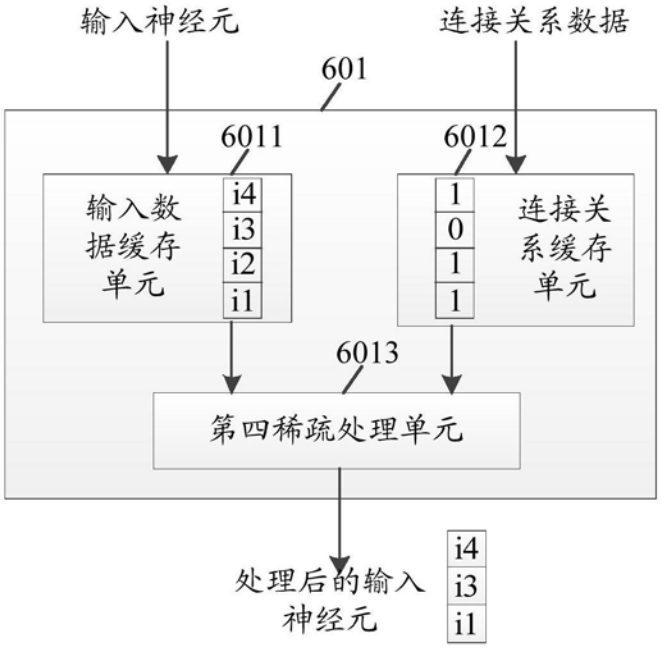

输入神经元
连接关系数据
601
6011
6012
输入数据缓存单元
i4
i3
i2
i1
1
0
1
1
连接关系缓存单元
6013
第四稀疏处理单元
处理后的输入神经元
i4
i3
i1

图8

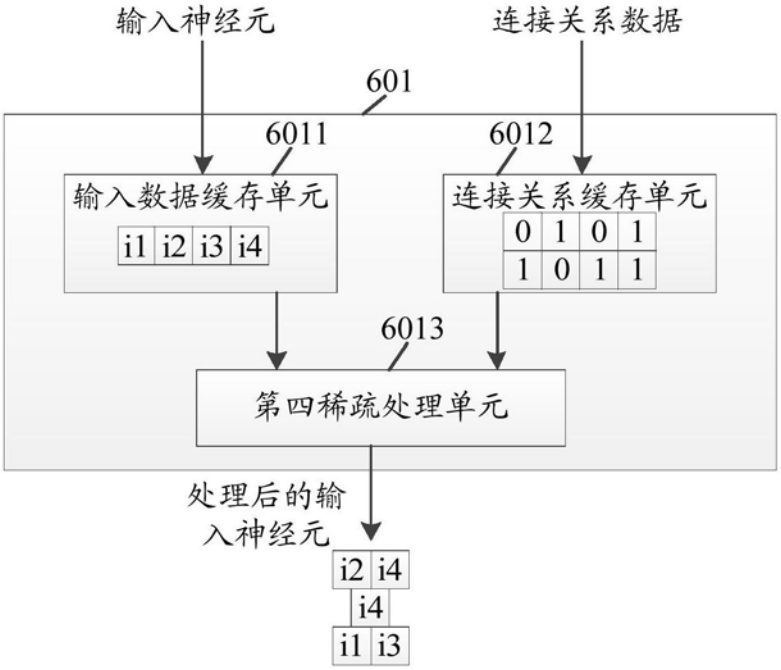

输入神经元
连接关系数据
601
6011
6012
输入数据缓存单元
i1 i2 i3 i4
连接关系缓存单元
0 1 0 1
1 0 1 1
6013
第四稀疏处理单元
处理后的输入神经元
i2 i4
i4
i1 i3

图9

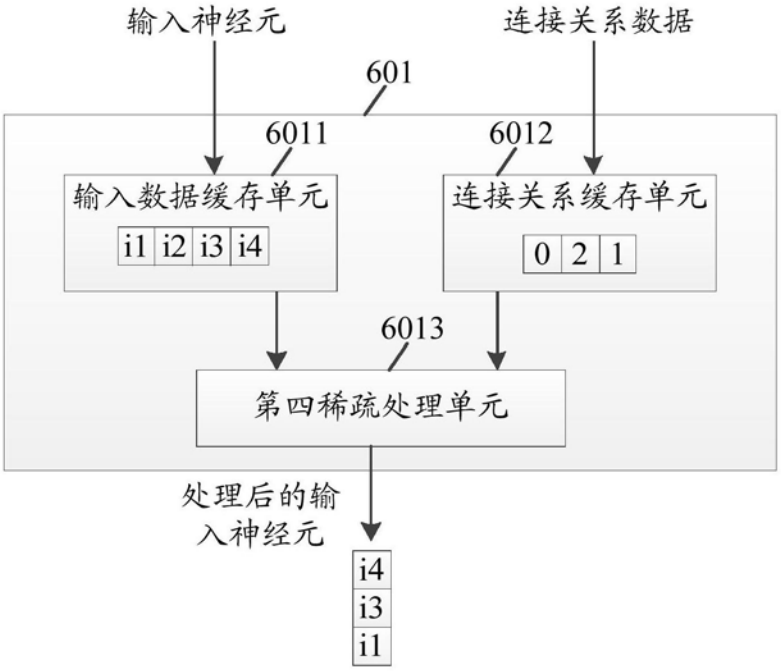

输入神经元
连接关系数据
601
6011
6012
输入数据缓存单元
i1 i2 i3 i4
连接关系缓存单元
0 2 1
6013
第四稀疏处理单元
处理后的输入神经元
i4
i3
i1

图10

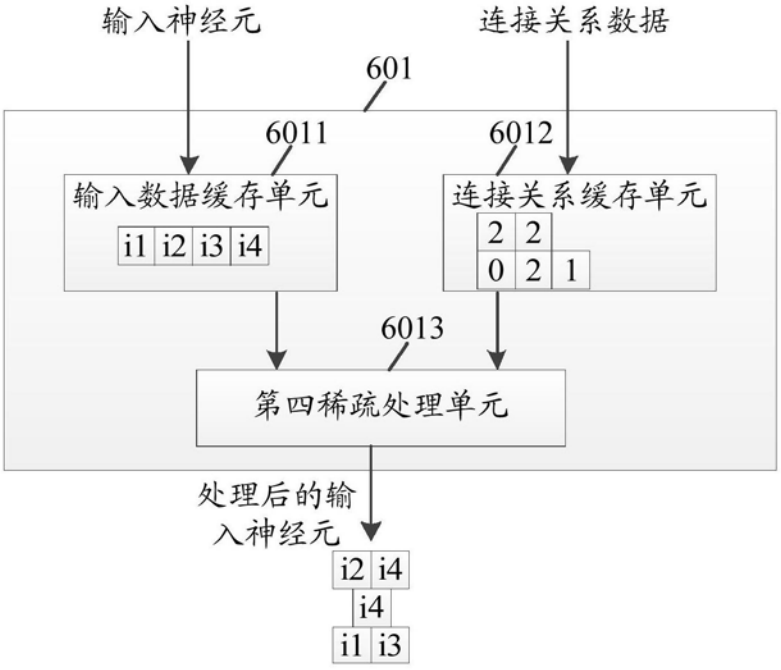

输入神经元
连接关系数据
601
6011
6012
输入数据缓存单元
i1 i2 i3 i4
连接关系缓存单元
2 2
0 2 1
6013
第四稀疏处理单元
处理后的输入神经元
i2 i4
i4
i1 i3

图11

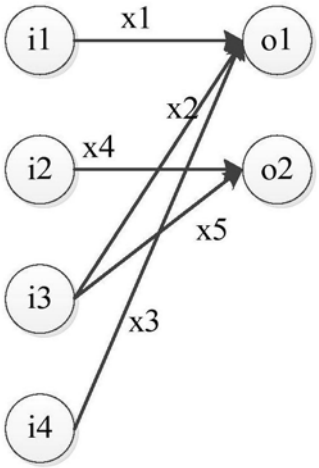

i1
i2
i3
i4
o1
o2
x1
x2
x3
x4
x5

图12

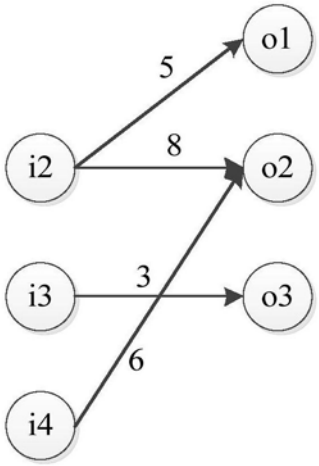

i2
i3
i4
o1
o2
o3
5
8
3
6

图13

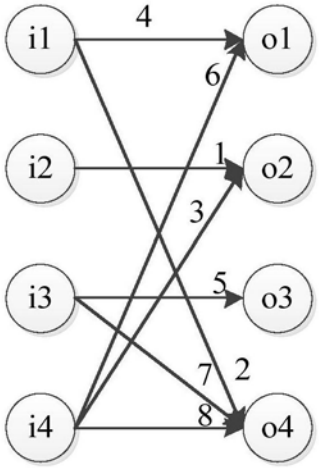

i1
i2
i3
i4
o1
o2
o3
o4
4
6
1
3
5
7
2
8

图14

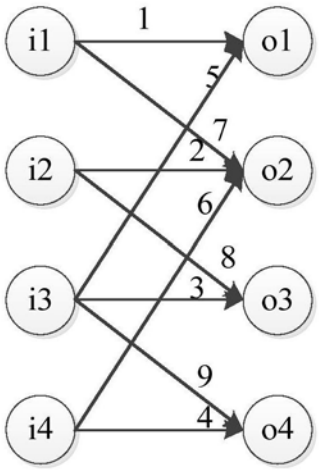


图15

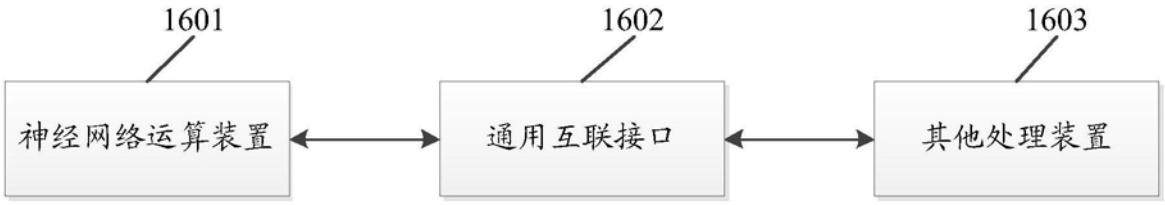


图16a

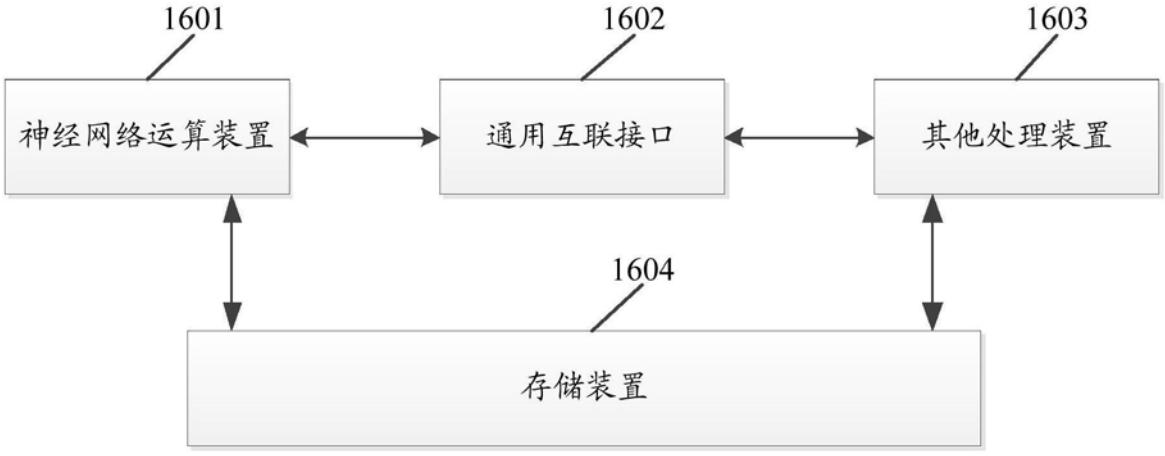


图16b

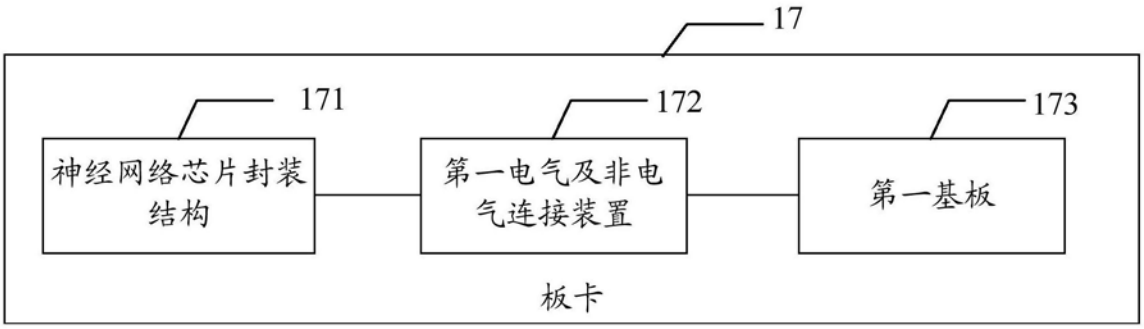


图17

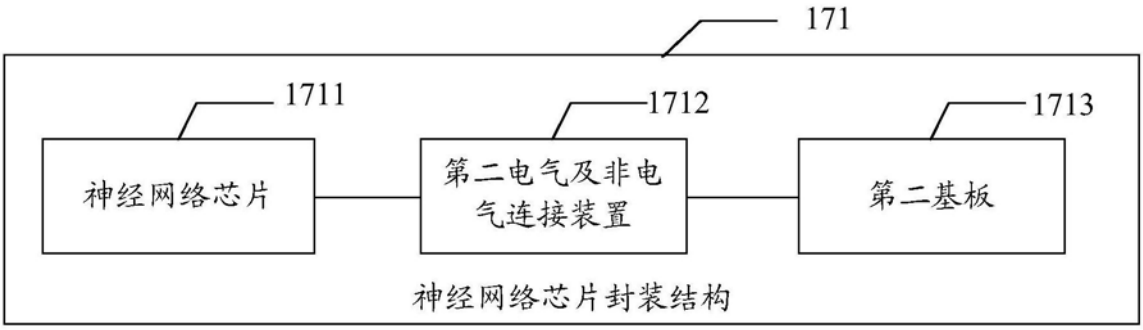


图18

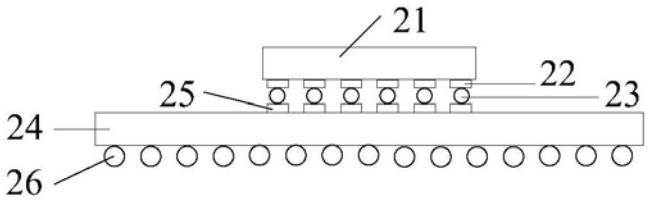


图19

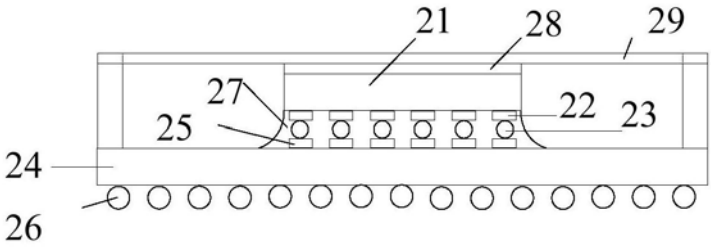


图20

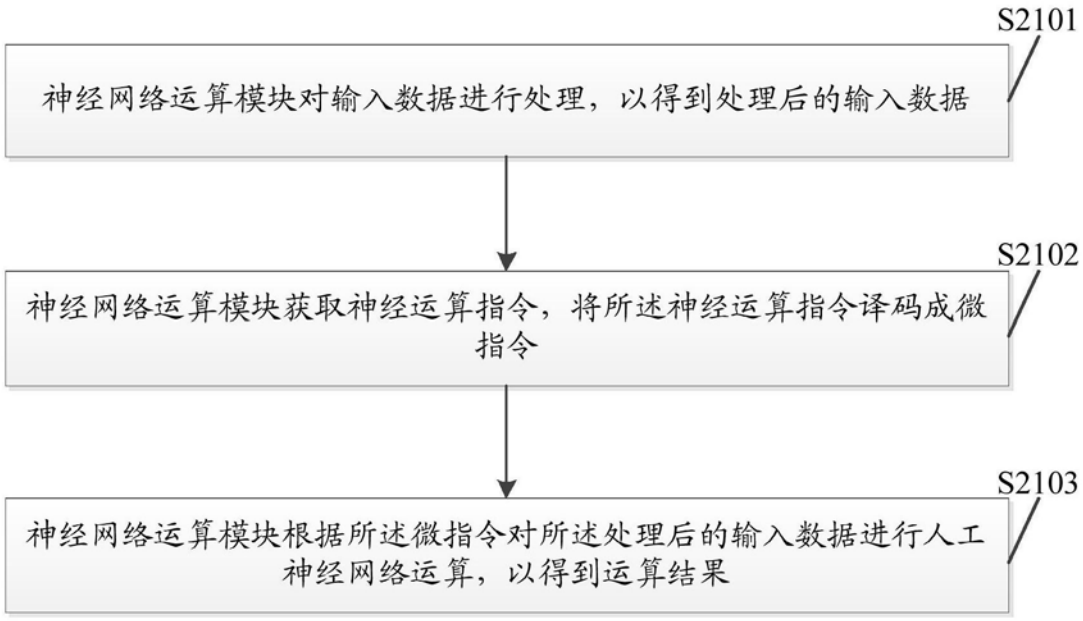

S2101
神经网络运算模块对输入数据进行处理，以得到处理后的输入数据
S2102
神经网络运算模块获取神经运算指令，将所述神经运算指令译码成微指令
S2103
神经网络运算模块根据所述微指令对所述处理后的输入数据进行人工神经网络运算，以得到运算结果

图21