(19)中华人民共和国国家知识产权局



(12)发明专利申请



(10)申请公布号 CN 111209244 A (43)申请公布日 2020.05.29

(21)申请号 201811392270.6

(22)申请日 2018.11.21

(71)申请人 上海寒武纪信息科技有限公司 地址 200120 上海市浦东新区南汇新城镇 环湖西二路888号

(72)发明人 不公告发明人

(74) 专利代理机构 北京华进京联知识产权代理 有限公司 11606

代理人 王程

(51) Int.CI.

GO6F 15/78(2006.01)

GO6N 3/063(2006.01)

GO6N 20/00(2019.01)

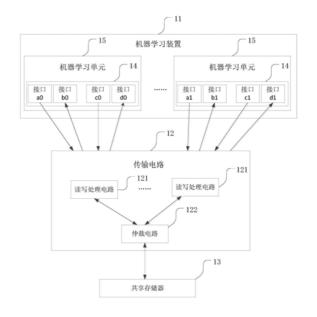
权利要求书2页 说明书13页 附图8页

(54)发明名称

数据处理装置及相关产品

(57)摘要

本发明涉及一种数据处理装置。该数据处理装置用于执行机器学习数据的处理,包括:机器学习装置、传输电路以及共享存储器,传输电路包括多个读写处理电路和一个仲裁电路,机器学习装置包括多个机器学习单元,每个机器学习单元包括至少一个运算单元,多个机器学习单元通过第一传输接口与传输电路连接,传输电路与共享存储器连接。数据处理装置通过仲裁电路对多个机器学习单元发送的数据操作信号进行仲裁,并根据仲裁结果从共享存储器中获取数据,使得数据处理装置在进行数据操作时,多个机器学习单元可以通过一个传输电路对共享存储器进行数据操作,并通过仲裁电路对多个数据操作信号数据操作,并通过仲裁电路对多个数据操作信号进行仲裁,在减小了硬件开销的同时,避免了信号的阻塞。



1.一种数据处理装置,其特征在于,所述数据处理装置用于执行机器学习数据的处理, 所述数据处理装置包括:机器学习装置、传输电路以及共享存储器,所述传输电路包括多个 读写处理电路和一个仲裁电路,所述机器学习装置包括多个机器学习单元,每个机器学习 单元包括至少一个运算单元,所述多个机器学习单元通过第一传输接口与所述传输电路连 接,所述传输电路与所述共享存储器连接;

所述仲裁电路,用于对所述多个机器学习单元发送的数据操作信号进行仲裁,并根据仲裁成功的数据操作信号从所述共享存储器中获取所述机器学习装置所需的输入数据;

所述读写处理电路,用于根据所述仲裁成功的数据操作信号所携带的地址信息或者所述数据操作信号的类型,从所述多个机器学习单元中确定目标机器学习单元或者目标运算单元;以及,将所述输入数据返回至所述目标机器学习单元或者目标运算单元。

- 2.根据权利要求1所述的数据处理装置,其特征在于,所述机器学习装置用于根据所述输入数据,执行机器学习运算,得到输出数据。
- 3.根据权利要求2所述的数据处理装置,其特征在于,所述机器学习装置还用于将所述 输出数据通过所述传输电路传输至所述共享存储器进行数据存储。
- 4.根据权利要求1所述的数据处理装置,其特征在于,所述读写处理电路包括以下处理 电路中的任一种:单播读处理电路、广播处理电路;所述数据操作信号包括以下至少一种: 单播读请求、单播读指令、多播指令、广播指令;

其中,单播类型的处理电路用于处理单播类型的信号,广播类型的处理电路用于处理 多播或者广播类型的信号。

- 5.根据权利要求4所述的数据处理装置,其特征在于,所述仲裁电路,具体用于确定多个读写处理电路发送的数据操作信号的优先级,并将优先级最高的数据操作信号作为仲裁成功的数据操作信号。
- 6.根据权利要求4所述的数据处理装置,其特征在于,所述仲裁电路,具体用于在多个读写处理电路发送的数据操作信号优先级相同时,根据所述多个数据操作信号的类型以及预设的执行条件,确定仲裁成功的数据操作信号。
- 7.根据权利要求6所述的数据处理装置,其特征在于,若所述数据操作信号为单播类型的信号,则所述执行条件包括:发送所述单播类型的信号的机器学习单元的通道空闲,或者发送所述单播类型的信号的机器学习单元中的运算单元的通道空闲。
- 8.根据权利要求6所述的数据处理装置,其特征在于,若所述数据操作信号为多播类型的信号,则所述执行条件包括:发送所述多播类型的信号的机器学习单元的通道空闲、且所述多播类型的信号指定的目标机器学习单元的通道空闲;或者,发送所述多播类型的信号的机器学习单元中的运算单元的通道空闲、且所述多播类型的信号指定的目标运算单元的通道空闲。
- 9.根据权利要求6所述的数据处理装置,其特征在于,若所述数据操作信号为广播类型的信号,则所述执行条件包括:发送所述广播类型的信号的机器学习单元的通道空闲、且其他剩余的机器学习单元的通道空闲;或者,发送所述广播类型的信号的机器学习单元中的运算单元的通道空闲、且其他剩余的机器学习单元中的运算单元的通道空闲。
- 10.根据权利要求9所述的数据处理装置,其特征在于,所述传输电路还包括第二传输接口,所述第二传输接口中的各个接口与所述第一传输接口中的各个接口一对应连接,

- 且一个机器学习单元与一个读写处理电路对应连接。
- 11.根据权利要求10所述的数据处理装置,其特征在于,所述一个机器学习单元中的多个运算单元共享所述第一传输接口中的一个发送接口,且每个运算单元对应一个数据接收接口。
- 12.根据权利要求10所述的数据处理装置,其特征在于,所述一个机器学习单元中的多个运算单元各自对应所述第一传输接口中的一个发送接口和一个数据接收接口。
- 13.根据权利要求9所述的数据处理装置,其特征在于,所述传输电路还包括第二传输接口,所述多个机器学习单元共享所述第二传输接口中的一个信号接收接口以及一个数据返回接口。
- 14.根据权利要求9所述的数据处理装置,其特征在于,所述读写处理电路还包括信号队列,所述信号队列用于存储各个机器学习单元发送的数据操作信号;

所述读写处理电路,还用于在接收到所述数据操作信号时,判断所述请求队列是否存在剩余空间,若是,则将所述数据操作信号缓存值所述请求队列中;若否,则阻塞所述数据操作信号。

15.根据权利要求14所述的数据处理装置,其特征在于,若读写处理电路为广播处理电路,则所述信号队列包括指令队列和请求队列;

所述指令队列用于缓存所述广播处理电路接收到的指令类型的信号;

所述请求队列用于缓存解析所述指令类型的信号后所得到的请求类型的信号。

16.根据权利要求9所述的数据处理装置,其特征在于,所述机器学习单元还包括:与所述运算单元连接的控制器单元;所述运算单元包括一个主处理电路和多个从处理电路,所述运算单元通过所述第一传输接口与所述传输电路连接;

所述控制器单元,用于通过所述第一传输接口中的发送接口向所述传输电路发送所述数据操作信号和所述输出数据,并通过所述第一传输接口中的接收接口接收所述传输电路从所述共享存储器获取的所述输入神经元数据和所述权值数据,并将所述输入神经元数据和所述权值数据发送给所述主处理电路和/或所述从处理电路;

所述主处理电路,用于将所述输入数据分发至所述多个从处理电路;所述多个从处理 电路,用于依据所述主处理电路传输的神经元数据和权值数据并行执行中间运算得到多个 中间结果,并将多个中间结果传输给所述主处理电路;

所述主处理电路,还用于对所述多个中间结果执行后续处理得到计算结果。

- 17.根据权利要求1-16中任一项所述的数据处理装置,其特征在于,所述输入数据包括输入数据,所述输出数据包括输出数据。
- 18.一种组合处理装置,其特征在于,所述组合处理装置包括如权利要求1-17中任一项 所述的数据处理装置、通用互联接口和除所述数据处理装置以外的其它处理装置;所述数 据处理装置与所述其它处理装置进行交互。
- 19.一种机器学习芯片,其特征在于,所述机器学习芯片包括如权利要求17所述的组合处理装置和/或1-17中任一项所述的组合处理装置。
 - 20.一种板卡,其特征在于,所述板卡包括如权利要求19所述的机器学习芯片。
 - 21.一种电子设备,其特征在于,所述电子设备包括如权利要求20所述的板卡。

数据处理装置及相关产品

技术领域

[0001] 本申请涉及信息处理技术领域,特别是涉及一种数据处理装置及相关产品。

背景技术

[0002] 随着目前人工神经网络的不断发展,越来越多的多架构的机器学习芯片逐渐问世,这些机器学习芯片在访问或者处理存储器中的数据时,由于机器学习芯片所需的数据量大,且机器学习芯片对数据处理的速度要求高,在进行数据访问或操作的过程中,常常通过增加硬件数量的方式来提升数据传输的带宽,进而提升数据处理的速度,以满足机器学习芯片对数据处理的速度要求高的特点。

[0003] 然而采用上述方法,机器学习芯片进行数据访问或操作时,会导致硬件开销大,硬件冗余。

发明内容

[0004] 本申请实施例提供了一种数据处理装置及相关产品,可以减少机器学习芯片在数据访问时的硬件开销,降低了机器学习芯片在进行数据操作时的硬件冗余。

[0005] 第一方面,一种数据处理装置,所述数据处理装置用于执行机器学习数据的处理, 所述数据处理装置包括:机器学习装置、传输电路以及共享存储器,所述传输电路包括多个 读写处理电路和一个仲裁电路,所述机器学习装置包括多个机器学习单元,每个机器学习 单元包括至少一个运算单元,所述多个机器学习单元通过第一传输接口与所述传输电路连 接,所述传输电路与所述共享存储器连接;

[0006] 所述仲裁电路,用于对所述多个机器学习单元发送的数据操作信号进行仲裁,并根据仲裁成功的数据操作信号从所述共享存储器中获取所述机器学习装置所需的输入数据:

[0007] 所述读写处理电路,用于根据所述仲裁成功的数据操作信号所携带的地址信息或者所述数据操作信号的类型,从所述多个机器学习单元中确定目标机器学习单元或者目标运算单元;以及,将所述输入数据返回至所述目标机器学习单元或者目标运算单元。

[0008] 第二方面,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数据处理装置、通用互联接口,和其他处理装置。该数据处理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。该组合处理装置还可以包括存储装置,该存储装置分别与所述数据处理装置和所述其他处理装置连接,用于保存所述数据处理装置和所述其他处理装置的数据。

[0009] 第三方面,本申请实施例提供了一种机器学习芯片,该机器学习芯片包括上述第一方面所述的数据处理装置、上述第二方面所述的组合处理装置。

[0010] 第四方面,本申请实施例提供了一种机器学习芯片封装结构,该机器学习芯片封装结构包括上述第三方面所述的机器学习芯片;

[0011] 第五方面,本申请实施例提供了一种板卡,该板卡包括上述第四方面所述的机器

学习芯片封装结构。

[0012] 第六方面,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三方面所述的机器学习芯片或者上述第五方面所述的板卡。

[0013] 上述数据处理装置、方法及相关产品,数据处理装置通过仲裁电路对多个机器学习单元发送的数据操作信号进行仲裁,并根据仲裁结果从共享存储器中获取机器学习装置所需的输入数据,使得数据处理装置在进行数据操作时,多个机器学习单元可以通过一个传输电路对共享存储器进行数据操作,并通过仲裁电路对多个数据操作信号进行仲裁,在减小了硬件开销的同时,避免了多个数据操作信号的阻塞。

附图说明

[0014] 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。

[0015] 图1为一个实施例中数据处理装置的结构示意图;

[0016] 图2为一个实施例中机器学习单元的结构示意图;

[0017] 图3为一个实施例中数据处理装置的结构示意图;

[0018] 图4为一个实施例中数据处理装置的结构示意图:

[0019] 图5为一个实施例中数据处理装置的结构示意图:

[0020] 图6为一个实施例中数据处理装置的结构示意图:

[0021] 图7为一个实施例中一种组合处理装置的结构示意图;

[0022] 图8为一个实施例中另一种组合处理装置的结构示意图;

[0023] 图9为一个实施例中一种板卡的结构示意图。

具体实施方式

[0024] 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。

[0025] 本申请的说明书和权利要求书及附图中的术语"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。

[0026] 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。

[0027] 本发明实施例涉及的数据处理装置,可以通过软件、硬件或者软硬件结合的方式

实现;该数据处理装置可以为图1所示的部分或者全部。该数据处理装置用于执行机器学习数据的处理,数据处理装置可以包括:机器学习装置11、传输电路12以及共享存储器13,传输电路12包括多个读写处理电路121和一个仲裁电路122,仲裁电路122用于对多个机器学习单元15发送的数据操作信号进行仲裁,并根据仲裁成功的数据操作信号从共享存储器13中获取机器学习装置11所需的输入数据;读写处理电路121用于根据仲裁成功的数据操作信号所携带的地址信息或者数据操作信号的类型,从多个机器学习单元中确定目标机器学习单元或者目标运算单元;以及,将输入数据返回至目标机器学习单元或者目标运算单元;机器学习装置11包括多个机器学习单元15,每个机器学习单元15包括至少一个运算单元151,多个机器学习单元通过第一传输接口14与传输电路12连接,传输电路12与共享存储器13连接。

[0028] 可选的,所述机器学习装置11可以用于根据输入数据,执行机器学习运算,得到输出数据。可选的,所述机器学习装置11还可以用于将输出数据通过传输电路12传输至共享存储器13进行数据存储。具体的,若机器学习装置11用于执行一种神经网络运算时,机器学习装置11可以根据输入神经元数据和权值数据,执行人工神经网络运算,得到输出神经元数据,以及将输出神经元数据作为新的输入神经元数据,并通过传输电路12传输至共享存储器13进行数据存储。

[0029] 需要说明的是,上述机器学习单元、传输电路、共享存储器以及各类接口均可以通过硬件电路的方式实现。示例性地,传输电路可以是广播总线(broadcast bus);共享存储器可以是非易失性和/或易失性存储器,包括但不限于随机存取存储器(RAM)、高速缓冲存储器等;各类接口可以对应一个或多个数据I/0(in/out,读入读出)接口或I/0引脚。

[0030] 本申请提出的数据处理装置可以应用于机器学习运算,其中,机器学习运算包括神经网络运算、k-means运算、支持向量机运算等。当上述机器学习装置进行神经网络计算时,可选地,上述输入数据可以包括输入神经元数据和权值数据,该输入神经元数据和权值数据是机器学习装置在执行一种人工神经网络运算时需要输入的数据;相应的,上述输出数据可以包括输出神经元数据,该输出神经元数据是机器学习装置在执行一种人工神经网络运算时输出的中间结果或最终结果。

[0031] 以神经网络运算为例(如果不进行额外说明,本实施例中均以神经网络运算为例进行说明),本申请提出的数据处理装置可以进行神经网络中的一层的运算。对于多层神经网络,其实现过程可以是:在正向运算中,当上一层人工神经网络执行完成之后,下一层的运算指令会将运算单元中计算出的输出神经元数据作为下一层的输入神经元数据进行运算(或者是对该输出神经元数据进行某些操作再作为下一层的输入神经元数据),同时,将权值数据也替换为下一层的权值数据;在反向运算中,当上一层人工神经网络的反向运算执行完成后,下一层运算指令会将运算单元中计算出的输入神经元梯度(同样可以作为一种输入的神经元数据)作为下一层的输出神经元梯度进行运算(同样可以作为一种输出的神经元数据)(或者是对该输入神经元梯度进行某些操作再作为下一层的输出神经元梯度),同时将权值数据替换为下一层的权值数据。

[0032] 参照图1所示,在一种可选的方案中,上述机器学习装置11可以包括多个机器学习单元15。针对多层神经网络的运算,以正向运算中某一层神经网络的计算为例进行说明。在一种实施方式中,上述机器学习装置可以通过多个机器学习单元(MLU,Machine Learning

Unit)并行计算神经网络中该层所有神经元的输出神经元数据。示例性地,该机器学习装置包括4个机器学习单元,该层神经网络具有100个神经元,则可以分配由每个机器学习单元处理其中25个神经元,可以通过设置对应的运算指令来实现;在该过程中,每一个机器学习单元均可以通过传输电路从共享存储器中获取分配的该层25个神经元分别对应的输入神经元数据和权值数据,以此计算得到分配的该层25个神经元的输出神经元数据,并可以将分配的该层25个神经元的输出神经元数据通过传输电路传输至共享存储器进行存储。可以理解的是,上述各个机器学习单元在处理其分配的该层的多个神经元数据时,可以是并行计算处理的;如此进行一层又一层神经网络的并行计算,可以实现神经网络计算的并行处理,提高了处理效率。

[0033] 当多个机器学习单元15同时通过第一传输接口14向传输电路12发送数据操作信号时,可以通过第一传输接口14向读写处理电路121发送数据操作信号,读写处理电路121可以是一个读写处理电路,也可以是多个读写处理电路,当读写处理电路121为多个读写处理电路时,可以是一个机器学习单元15对应一个读写处理电路,也可以是一个机器学习单元15对应多个读写处理电路。读写处理电路121将数据操作信号发送给仲裁电路122,仲裁电路122对多个数据操作信号进行仲裁,并根据仲裁成功的数据操作信号,从共享存储器13中获取数据操作信号对应的机器学习单元需要的输入神经元数据和权值数据。读写处理电路121可以根据数据操作信号所携带的地址信息,或者根据数据操作信号的类型,确定出目标机器学习单元,或目标运算单元,或目标运算单元,并将输入神经元数据和权值数据返回目标机器学习单元,或目标运算单元。

[0034] 示例性的,当机器学习装置包括4个机器学习单元,分别为机器学习单元0、机器学习单元1、机器学习单元2和机器学习单元3,其对应了4个读写处理电路,分别为读写处理电路0、读写处理电路1、读写处理电路2和读写处理电路3。其中,机器学习单元0、机器学习单元1、机器学习单元2和机器学习单元3通过第一传输接口14分别向读写处理电路0、读写处理电路1、读写处理电路2和读写处理电路3发送数据操作信号,其可以是向读写处理电路0发送数据操作信号0、向读写处理电路1发送数据操作信号1、向读写处理电路2发送数据操作信号2和向读写处理电路3发送数据操作信号3,读写处理电路0、读写处理电路1、读写处理电路2和读写处理电路3分别将数据操作信号3、数据操作信号1、数据操作信号2和数据操作信号3发送给仲裁电路122进行仲裁,仲裁电路122对多个数据操作信号进行仲裁,确定数据操作信号2为仲裁成功的数据操作信号,根据数据操作信号2,从共享存储器13中获取输入神经元数据和权值数据;读写处理电路2根据数据操作信号2中携带的地址信息,地址信息包括机器学习单元1和机器学习单元2的地址,确定目标机器学习单元为机器学习单元1和机器学习单元2

[0035] 在另一种可选的方案中,上述机器学习装置还可以利用多个机器学习单元按照一定先后顺序分别计算神经网络中各层所有神经元的输出神经元数据。在该过程中,前一个机器学习单元可以将该层所有神经元的输出神经元数据通过传输电路传输至共享存储器进行存储,以供后一个机器学习单元可以提取到该层所有神经元的输出神经元数据作为下一层的输入神经元数据进行计算。可以理解的是,上述应用适用于每一层的神经网络计算量不大的场景,例如每一层的神经元数目较少的神经网络的计算。

[0036] 参照图2所示,这里对机器学习单元15进行详细说明。在一种方案中,机器学习单元15可以包括:至少一个运算单元151、以及与运算单元151连接的控制器单元152;运算单元151可以包括一个主处理电路151a和多个从处理电路151b,运算单元151通过第一传输接口14与传输电路12连接;

[0037] 控制器单元152,用于通过第一传输接口14向传输电路12发送数据操作信号和输出神经元数据,并通过第一传输接口14接收传输电路12从共享存储器13获取的输入神经元数据和权值数据,并将输入神经元数据和权值数据发送给主处理电路151a和/或从处理电路151b:

[0038] 主处理电路151a,用于将输入神经元数据和权值数据分发至多个从处理电路151b;多个从处理电路151b,用于依据神经元数据和权值数据并行执行中间运算得到多个中间结果,并将多个中间结果传输给主处理电路151a;主处理电路151a,还用于对多个中间结果执行后续处理得到计算结果,后续处理包括激活运算。具体地,该控制器单元152还可以获取计算指令,并解析该计算指令得到多个运算指令,将该多个运算指令发送给主处理电路。

[0039] 可以理解的是,在本实施例中,机器学习单元包括多个运算单元,各运算单元可以通过上述第一传输接口发送或接收数据。

[0040] 举例说明,在一个可选的技术方案中,主处理电路也可以包括一个控制器单元,该控制器单元可以包括主指令处理单元,具体用于将运算指令译码成微指令。当然在另一种可选方案中,从处理电路也可以包括另一个控制器单元,该另一个控制器单元包括从指令处理单元,具体用于接收并处理微指令。上述微指令可以为指令的下一级指令,该微指令可以通过对指令的拆分或解码后获得,能被进一步解码为各部件、各单元或各处理电路的控制信号。例如,乘积微指令为卷积指令的下一级指令。

[0041] 示例型性的,以上述的机器学习单元的结构为例,对该机器学习单元的神经网络运算流程进行详细说明,参见以下步骤S101-S106:

[0042] S101,在控制器单元的指令存储单元的首地址处预先存入一条I0指令;

[0043] S102,控制器单元从指令存储单元的首地址读取该I0指令,然后根据该I0指令译出的控制信号,通过片外接口从片外存储器中获取该机器学习单元对应的神经网络运算指令,或者,通过传输电路从共享存储器中获取该机器学习单元对应的神经网络计算指令,并将获得的计算指令储存在指令存储单元中;

[0044] S103,控制器单元从指令存储单元中读入下一条I0指令,根据该I0指令译出的数据操作信号,通过传输电路从共享存储器中读取运算单元需要的所有数据块,包括需要的分配的该层神经元的输入神经元数据和权值数据,还可以包括用于作快速的激活函数运算的插值表、用于配置运算器件参数的常数表、偏置数据等;上述数据操作信号中包括上述数据块在共享存储器中的源地址;

[0045] S104,控制器单元从指令存储单元中读入下一条CONFIG(配置)指令,根据该CONFIG指令译出的控制信号,配置该层神经网络计算所需的各类常数,例如,运算单元根据激活函数所需的常数配置其内寄存器的值;

[0046] S105,控制器单元从指令存储单元中读入下一条COMPUTE(计算)指令,根据该COMPUTE指令译出的控制信号(即运算指令),运算单元将分配的该层神经元的输入神经元

数据、权值数据和运算指令传输给主处理电路;主处理电路可以将分配的该层神经元的输入神经元数据确定为广播数据,权值数据确定为分发数据,将一个分发数据分配成多个数据块,将多个数据块中的至少一个数据块、广播数据以及多个运算指令中的至少一个运算指令发送给从处理电路,由从处理电路根据乘法处理电路、累加处理电路等得到中间结果,并由主处理电路根据中间结果和激活处理电路等得到分配的该层神经元输出的神经元数据;

[0047] S106,控制器单元从指令存储单元中读入下一条I0指令,根据该I0指令译出的数据操作信号,将上述输出的神经元数据,通过传输电路传输至共享存储器中存储,下一层部分神经元的输入神经元数据;上述数据操作信号中包括上述输出的神经元数据在共享存储器中的目的地址。

[0048] 以下示例性地对S105进行说明。以神经网络运算中的全连接运算为例,某层神经网络过程可以为:y=f(wx+b),其中,x为输入神经元矩阵,w为权值矩阵,b为偏置标量,f为激活函数,具体可以为:sigmoid函数,tanh、relu、softmax函数中的任意一个。这里假设主从处理电路之间为二叉树关系(一种树型关系),运算单元具有一个主处理电路和8个从处理电路,则上述S105的实现的方法可以为:控制器单元从共享存储器内获取输入神经元矩阵x,权值矩阵w以及全连接运算指令,将输入神经元矩阵x,权值矩阵w以及全连接运算指令传输给主处理电路;主处理电路确定该输入神经元矩阵x为广播数据,确定权值矩阵w为分发数据,将权值矩阵w拆分成8个子矩阵,然后将8个子矩阵通过树型模块分发给8个从处理电路,将输入神经元矩阵x广播给8个从处理电路,从处理电路并行执行8个子矩阵与输入神经元矩阵x的乘法运算和累加运算得到8个中间结果,将8个中间结果发送给主处理电路;主处理电路,用于将8个中间结果排序得到wx的运算结果,将该运算结果执行偏置b的运算后执行激活运算得到最终结果y。

[0049] 可以理解的是,因为上述各机器学习单元可以并行对某一层的各分配的神经元进行计算,因此共享存储器中可以存储有每一层所有神经元的输出神经元数据,以及下一层所有神经元需要的输入的神经元数据,而权值数据可以复用或从共享存储器中获取新一层的神经网络的权值数据。

[0050] 需要说明的是,各机器学习单元中可以包括一个运算单元,也可以包括多个运算单元,而且各运算单元的结构可以相同,也可以不同。其中,各运算单元的结构体现在主处理电路和各从处理电路的关系可以包括但不限于是树型、H型、脉冲阵列型的关系。本申请提供的技术方案将运算单元设置成一主多从结构,对于正向运算的计算指令,其可以将依据正向运算的计算指令将数据进行拆分,这样通过多个从处理电路即能够对计算量较大的部分进行并行运算,从而提高运算速度,节省运算时间,进而降低功耗。

[0051] 本实施例提供的用于执行机器学习数据处理的数据处理装置,包括:机器学习装置、传输电路以及共享存储器,传输电路包括多个读写处理电路和一个仲裁电路,机器学习装置包括多个机器学习单元,每个机器学习单元包括至少一个运算单元,多个机器学习单元通过第一传输接口与传输电路连接,传输电路与共享存储器连接。本实施例中,数据处理装置通过仲裁电路对多个机器学习单元发送的数据操作信号进行仲裁,并根据仲裁结果从共享存储器中获取机器学习装置所需的输入神经元数据和权值数据,使得数据处理装置在进行数据操作时,多个机器学习单元可以通过一个传输电路对共享存储器进行数据操作,

并通过仲裁电路对多个数据操作信号进行仲裁,在减小了硬件开销的同时,避免了多个数据操作信号的阻塞。

[0052] 在一个实施例中,继续如图1所示,读写处理电路包括以下处理电路中的任一种:单播读处理电路、广播处理电路;数据操作信号包括以下至少一种:单播读请求、单播读指令、多播指令、广播指令;其中,单播类型的处理电路用于处理单播类型的信号,广播类型的处理电路用于处理多播或者广播类型的信号。

[0053] 示例性地,单播读指令为某机器学习单元发送的对共享存储器中源地址的输入神经元数据和权值数据的读取指令,可以通过单播读指令向上述机器学习单元返回输入神经元数据和权值数据,该输入神经元数据和权值数据为该机器学习单元根据计算指令进行某一层分配的神经元计算过程中该分配的神经元所需的输入神经元数据和权值数据;广播指令为某机器学习单元发送的对共享存储器中源地址的输入神经元数据和权值数据的读取指令,可以通过广播指令向上述机器学习装置中的所有机器学习单元返回该输入神经元数据和权值数据,该输入神经元数据可以某一层所有的神经元所需的输入神经元数据,即上一层所有的输出神经元数据,该权值数据可以为复用的权值数据,例如卷积核;多播指令与广播指令的不同在于,多播指令的数据返回对象并非上述机器学习装置中的所有机器学习单元,而是与该多播指令中标记域对应的多个机器学习单元。此外,一般来说,指令和请求的区别在于,执行指令的开销比较大,但是指令包含的信息比较多;而执行请求开销比较小,但是请求包含的信息比较少。

[0054] 上述实施例中,数据处理装置通过仲裁电路对多个机器学习单元发送的数据操作信号进行仲裁,并根据仲裁结果从共享存储器中获取机器学习装置所需的输入神经元数据和权值数据,使得数据处理装置在进行数据操作时,多个机器学习单元可以通过一个传输电路对共享存储器进行数据操作,减小了硬件开销,避免了硬件冗余。通过下述实施例,详细说明仲裁模块如何确定多个读写处理电路发送的数据操作信号的优先级的具体过程。

[0055] 在一个实施例中,所述仲裁电路122,具体用于确定多个读写处理电路121发送的数据操作信号的优先级,并将优先级最高的数据操作信号作为仲裁成功的数据操作信号。

[0056] 其中,仲裁电路122可以按照预设的规则确定多个数据操作信号的优先级,从而使仲裁电路122可以根据各个数据操作信号的优先级来确定需要操作的对象,即仲裁成功的数据操作信号。其可以是根据各数据操作信号的发送时间作为仲裁依据,也可以是根据各数据操作信号中携带的传输速率信息来作为仲裁依据。例如,读写处理电路1发送数据操作信号的时间为T+1时刻,则根据数据操作信号的时间为T时刻,读写处理电路2发送数据操作信号的时间为T+1时刻,则根据数据操作信号发送的时间作为仲裁依据,得到读写处理电路1发送数据操作信号为优先级高的数据操作信号,即为仲裁成功的数据操作信号。仲裁电路122根据冲裁的结果,根据仲裁成功的读写处理电路1发送的数据操作信号,从共享存储器13中获取数据。

[0057] 本实施例所示的数据处理装置,通过仲裁电路确定多个读写处理电路发送的数据操作信号的优先级,并将优先级最高的数据操作信号作为仲裁成功的数据操作信号。在同时接收到多个数据操作信号时,通过仲裁电路,确定一个可以执行的数据操作信号,避免了同时执行多个数据操作信号而引起的数据阻塞的情况,进而使得多个机器学习单元可以通过一个传输电路对共享存储器进行数据操作,减小了硬件开销,避免了硬件冗余。

[0058] 在一个实施例中,仲裁电路122,具体用于在多个读写处理电路121发送的数据操

作信号优先级相同时,根据多个数据操作信号的类型以及预设的执行条件,确定仲裁成功的数据操作信号。

[0059] 其中,在上述实施例的基础上,当多个读写处理电路121发送的数据操作信号优先级相同时,仲裁电路122可以根据多个数据操作信号的类型以及预设的执行条件,确定仲裁成功的数据操作信号。

[0060] 其中,预设的执行条件可以是根据检测数据操作信号对应的数据传输通道是否空闲,来确定仲裁结果。若数据传输通道空闲,则仲裁该数据传输通道对应的数据操作信号为仲裁成功的数据操作信号;还可以是根据数据操作信号中携带的发送时间信息来确定仲裁结果。示例性的,若仲裁电路122接收到4个数据操作信号,分别为数据操作信号0、数据操作信号1、数据操作信号2和数据操作信号3,其中,数据操作信号1和数据操作信号2的优先级相同,其中,数据操作信号1为单播读指令,数据操作信号2为广播指令,根据数据操作信号1为单播读指令,数据操作信号2为广播指令,根据数据操作信号1的类型确定机器学习单元0、机器学习单元1为目标机器学习单元3为目标机器学习单元3为目标机器学习单元,此时,机器学习单元0、机器学习单元1和机器学习单元2的数据通道空闲,机器学习单元3的数据通道为忙碌,则仲裁电路122根据数据操作信号1为单播读指令,数据操作信号2为广播指令,及,机器学习单元3的数据通道为忙碌,则仲裁电路122根据数据操作信号1为单播读指令,数据操作信号2为广播指令,及,机器学习单元3的数据通道为忙碌,确定仲裁成功的数据操作信号为数据操作信号1。

[0061] 可选地,若数据操作信号为单播类型的信号,则上述执行条件包括:发送单播类型的信号的机器学习单元的通道空闲,或者发送单播类型的信号的机器学习单元中的运算单元的通道空闲。

[0062] 可选地,若数据操作信号为多播类型的信号,则上述执行条件包括:发送多播类型的信号的机器学习单元的通道空闲、且多播类型的信号指定的目标机器学习单元的通道空闲;或者,发送多播类型的信号的机器学习单元中的运算单元的通道空闲、且多播类型的信号指定的目标运算单元的通道空闲。

[0063] 可选地,若数据操作信号为广播类型的信号,则上述执行条件包括:发送广播类型的信号的机器学习单元的通道空闲、且其他剩余的机器学习单元的通道空闲;或者,发送广播类型的信号的机器学习单元中的运算单元的通道空闲、且其他剩余的机器学习单元中的运算单元的通道空闲。

[0064] 本实施例所示的数据处理装置,当多个读写处理电路发送的数据操作信号优先级相同时,仲裁电路可以根据多个数据操作信号的类型以及预设的执行条件,确定仲裁成功的数据操作信号。本实例中,当数据操作信号优先级相同时,通过数据操作信号的类型及预设的执行条件,确定仲裁成功的数据操作信号,进一步地避免同时执行多个数据操作信号而出现的数据阻塞的情况,进而使得多个机器学习单元可以通过一个传输电路对共享存储器进行数据操作,减小了硬件开销,避免了硬件冗余。

[0065] 在一个实施例中,参见图3所示,传输电路12还包括第二传输接口120,第二传输接口120中的各个接口与第一传输接口14中的各个接口一一对应连接,且一个机器学习单元15与一个读写处理电路121对应连接。

[0066] 其中,第一传输接口14可以将数据操作信号通过第二传输接口120发送给对应的读写处理电路121,传输电路12可以通过第二传输接口120将返回的机器学习装置所需的输

入神经元数据和权值数据返回第一传输接口14,再通过第一传输接口14返回目标机器学习单元或目标运算单元。第一传输接口14可以包括一个接口,也可以包括多个接口;第二传输接口120可以包括一个接口,也可以包括多个接口。示例性的,若第一传输接口14包括一个发送接口141和一个数据接收接口142,则第二传输接口120包括与一个发送接口141和一个返回接口142对应的第二接收接口1201和第二返回接口1202。

[0067] 可选地,参照图4所示,一个机器学习单元15中的多个运算单元151共享第一传输接口14中的一个发送接口141,且每个运算单元对应一个数据接收接口142。

[0068] 其中,当一个机器学习单元15包括多个运算单元151时,多个运算单元151可以共享第一传输接口14的一个发送接口141,一个机器学习单元15中的多个运算单元151通过共享的一个发送接口141将数据操作信号发送给传输电路12,传输电路12将获取的输入神经元数据和权值数据,通过目标运算单元151对应的数据接收接口142返回目标运算单元。

[0069] 因此,本实施例所示的数据处理装置,一个机器学习单元中的多个运算单元共享所述第一传输接口中的一个发送接口,且每个运算单元对应一个数据接收接口,能够有效节省机器学习单元中数据操作信号发送接口的数量,节省硬件资源,降低硬件的面积和功耗。

[0070] 在一个实施例中,参见图5所示,一个机器学习单元15中的多个运算单元151各自对应所述第一传输接口中的一个发送接口141和一个数据接收接口142。

[0071] 其中,如图5所示,一个运算单元151可以对应一个发送接口141和一个数据接收接口142,运算单元151通过与其对应的发送接口141将数据操作信号发送给传输电路12,传输电路12将获取的输入神经元数据和权值数据通过对应的数据接收接口142返回给对应的目标运算单元151。示例性的,运算单元1对应发送接口1和数据接收接口1,运算单元2对应发送接口2和数据接收接口2,运算单元1通过发送接口1将数据操作信号发送给传输电路12,传输电路12根据数据操作信号确定目标运算单元为运算单元1和运算单元2,则传输电路通过数据接收接口1和数据接收接口2将获取的输入神经元数据和权值数据返回给运算单元1和运算单元2。

[0072] 因此,本实施例所示的数据处理装置中,一个机器学习单元中的多个运算单元各自对应所述第一传输接口中的一个发送接口和一个数据接收接口,通过多个运算单元与第一传输接口中的发送接口和数据接收接口的一一对应,能够有效简化数据传输过程中的控制逻辑。

[0073] 在一个实施例中,参见图6所示,多个机器学习单元15共享第二传输接口120中的一个信号接收接口1201以及一个数据返回接口1202。

[0074] 其中,多个机器学习单元15可以共享第二传输接口120中的一个信号接收接口1201以及一个数据返回接口1202。示例性的,若读写处理电路121为广播读处理电路,则多个机器学习单元发送的数据操作信号,通过一个信号接收接口1201发送给广播读处理电路,广播读处理电路根据数据操作信号,获取输入神经元数据和权值数据,并根据数据操作信号中的地址信息,将输入神经元数据和权值数据通过数据返回接口1202返回给目标机器学习单元。

[0075] 本实施例提供的数据处理装置,多个机器学习单元共享第二传输接口中的一个信号接收接口,以及一个数据返回接口。本实施例中,数据处理装置通过共享第二传输接口中

的一个信号接收接口,以及一个数据返回接口,进一步地减小了硬件开销,避免了硬件冗余。

[0076] 在一种实施例中,继续参见图6所示,读写处理电路121还包括信号队列,信号队列用于存储各个机器学习单元15发送的数据操作信号;读写处理电路121,还用于在接收到数据操作信号时,判断信号队列是否存在剩余空间,若是,则将数据操作信号缓存至信号队列中;若否,则阻塞数据操作信号。

[0077] 其中,信号队列可以用于存储各个机器学习单元15发送的数据操作信号,其可以设置在读写处理电路121之外,也可以是设置在读写处理电路121内。当读写处理电路121接收到了数据操作信号时,可以通过向信号队列发送内存查询指令,获取信号队列的存储空间,若信号队列的存储空间的大小能够存储数据操作信号,则将数据操作信号缓存至信号队列中;若信号队列的存储空间的大小不能够存储数据操作信号,则阻塞数据操作信号。

[0078] 本实施例所示的数据处理装置,读写处理电路还包括信号队列,信号队列用于存储各个机器学习单元发送的数据操作信号;读写处理电路,还用于在接收到数据操作信号时,判断信号队列是否存在剩余空间,若是,则将数据操作信号缓存至信号队列中;若否,则阻塞数据操作信号。本实施例中,当读写处理电路接收到多个数据操作信号时,可以通过将数据操作信号缓存至信号队列,或阻塞数据操作信号,使得数据操作信号可以一个一个的发送个仲裁电路进行处理,避免了数据操作信号堵塞的情况,进而使得多个机器学习单元可以通过一个传输电路对共享存储器进行数据操作,减小了硬件开销,避免了硬件冗余。

[0079] 可选地,若读写处理电路121为广播处理电路,则信号队列包括指令队列和请求队列;指令队列用于缓存广播处理电路接收到的指令类型的信号;请求队列用于缓存解析指令类型的信号后所得到的请求类型的信号。

[0080] 其中,当读写处理电路121为广播处理电路时,信号队列可以包括指令队列和请求队列,其可以是将接收到的各个机器学习单元15发送的指令类型的信号存储在指令队列,通过广播处理电路对指令类型的信号进行解析处理,以获得请求类型的信号,并将获得的请求类型的信号存储在请求队列中。其中,指令队列用于缓存广播处理电路接收到的指令类型的信号;请求队列用于缓存解析指令类型的信号后所得到的请求类型的信号。

[0081] 本实施例所示的数据处理装置,当读写处理电路为广播处理电路,信号队列包括指令队列和请求队列,指令队列用于缓存广播处理电路接收到的指令类型的信号;请求队列用于缓存解析指令类型的信号后所得到的请求类型的信号。本实施例中,通过将指令类型的信号和请求类型的信号和请求类型的信号分别存储在指令队列和请求队列中,使得指令类型的信号和请求类型的信号可以分别一个一个的发送个仲裁电路进行处理,进一步地避免了数据操作信号堵塞的情况,进而使得多个机器学习单元可以通过一个传输电路对共享存储器进行数据操作,减小了硬件开销,避免了硬件冗余。

[0082] 参照图7所示,本申请还提供了一种组合处理装置,其包括上述数据处理装置、通用互联接口和除上述数据处理装置以外的其它处理装置;数据处理装置与其它处理装置进行交互,共同完成用户指定的计算操作。其中,其他处理装置,包括中央处理器CPU、图形处理器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。其他处理装置所包括的处理器数量不做限制。其他处理装置作为数据处理装置与外部数据和控制的接口,包括数据搬运,完成对本数据处理装置的开启、停止等基本控制;其他处理装置也可以

和数据处理装置协作共同完成运算任务。通用互联接口,用于在所述数据处理装置与其他处理装置间传输数据和控制指令。该数据处理装置从其他处理装置中获取所需的输入数据,写入数据处理装置片上的共享存储器;可以从其他处理装置中获取控制指令,写入数据处理装置片上的机器学习装置;也可以读取数据处理装置的共享存储器中的数据并传输给其他处理装置。

[0083] 可选的,参照图8所示,上述组合处理装置还可以包括存储装置,存储装置分别与所述数据处理装置和所述其他处理装置连接。存储装置用于保存在所述数据处理装置和所述其他处理装置的数据,尤其适用于所需要运算的数据在本数据处理装置或其他处理装置的内部存储中无法全部保存的数据。

[0084] 该组合处理装置可以作为手机、机器人、无人机、视频监控设备等设备的SOC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。此情况时,该组合处理装置的通用互联接口与设备的某些部件相连接。某些部件譬如摄像头,显示器,鼠标,键盘,网卡,wifi接口。

[0085] 在一个实施例中,本申请还提供了一种机器学习芯片,其包括了上述数据处理装置和/或组合处理装置。

[0086] 在一个实施例中,本申请还提供了一种机器学习芯片封装结构,其包括了上述机器学习芯片。

[0087] 在一个实施例中,本申请还提供了一种板卡,其包括了上述机器学习芯片封装结构。参照图9所示,上述板卡除了包括上述芯片封装结构81以外,还可以包括其他的配套部件,该配套部件包括但不限于:存储器件82、接口装置83和控制器件84;所述存储器件82与所述芯片封装结构81内的机器学习芯片811通过总线连接,用于存储数据,所述存储器件82可以包括多组存储单元821。每一组所述存储单元821与所述机器学习芯片811通过总线连接。可以理解,每一组所述存储单元821可以是DDR SDRAM (Double Data Rate SDRAM,双倍速率同步动态随机存储器)。

[0088] DDR不需要提高时钟频率就能加倍提高SDRAM的速度。DDR允许在时钟脉冲的上升沿和下降沿读出数据。DDR的速度是标准SDRAM的两倍。在一个实施例中,所述存储装置可以包括4组所述存储单元。每一组所述存储单元可以包括多个DDR4颗粒(芯片)。在一个实施例中,所述机器学习芯片内部可以包括4个72位DDR4控制器,上述72位DDR4控制器中64bit用于传输数据,8bit用于ECC校验。在一个实施例中,每一组所述存储单元包括多个并联设置的双倍速率同步动态随机存储器。DDR在一个时钟周期内可以传输两次数据。在所述芯片中设置控制DDR的控制器,用于对每个所述存储单元的数据传输与数据存储的控制。

[0089] 所述接口装置83与所述芯片封装结构81内的机器学习芯片811电连接。所述接口装置83用于实现所述机器学习芯片811与外部设备(例如服务器或计算机)之间的数据传输。例如在一个实施例中,所述接口装置83可以为标准PCIE(peripheral component interconnect express,一种高速串行计算机扩展总线标准)接口。比如,待处理的数据由服务器通过标准PCIE接口传递至所述机器学习芯片,实现数据转移。优选的,当采用PCIE 3.0X 16接口传输时,理论带宽可达到16000MB/s。在另一个实施例中,所述接口装置83还可以是其他的接口,本申请并不限制上述其他的接口的具体表现形式,所述接口装置能够实现转接功能即可。另外,所述机器学习芯片811的计算结果仍由所述接口装置83传送回外部

设备(例如服务器)。

[0090] 所述控制器件84与所述机器学习芯片811电连接。所述控制器件84用于对所述芯片的状态进行监控。具体的,所述机器学习芯片811与所述控制器件84可以通过SPI (Serial Peripheral Interface,串行外设接口)接口电连接。所述控制器件可以包括单片机 (Micro Controller Unit,MCU)。如所述机器学习芯片可以包括多个数据处理装置和/或组合处理装置,可以带动多个负载。因此,所述机器学习芯片可以处于多负载和轻负载等不同的工作状态。通过所述控制器件84可以实现对所述机器学习芯片中多个数据处理装置和/或组合处理装置的工作状态的调控。

[0091] 在一些实施例里,申请了一种电子设备,其包括了上述板卡。电子设备包括数据处理装置、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。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。

[0092] 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。

[0093] 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。

[0094] 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。

[0095] 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。

[0096] 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。

[0097] 本领域普通技术人员可以理解上述实施例的全部或部分处理过程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。

[0098] 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。

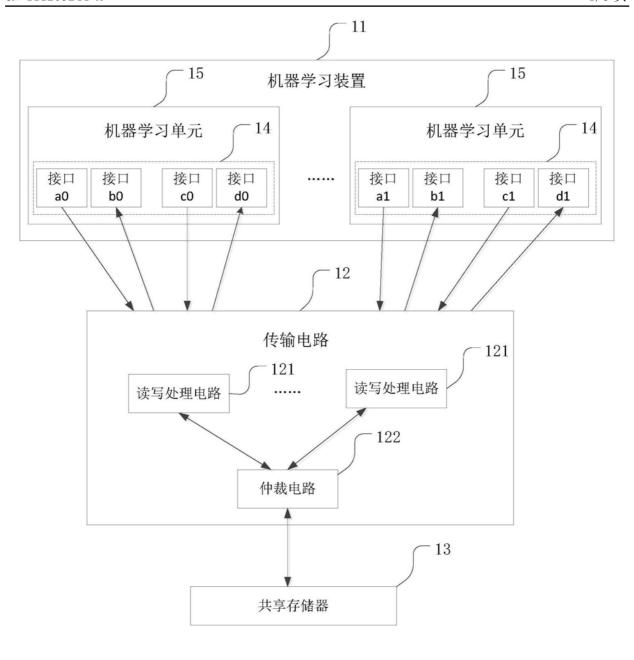


图1

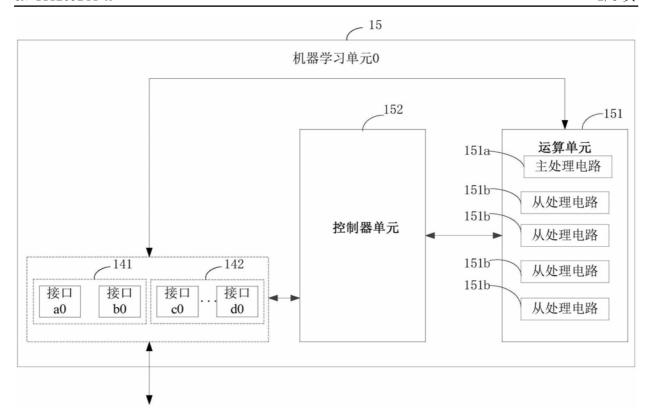


图2

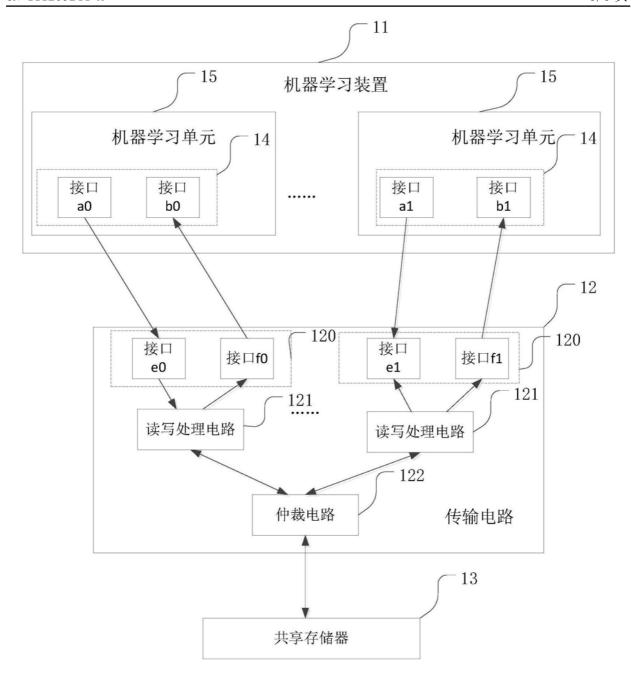


图3

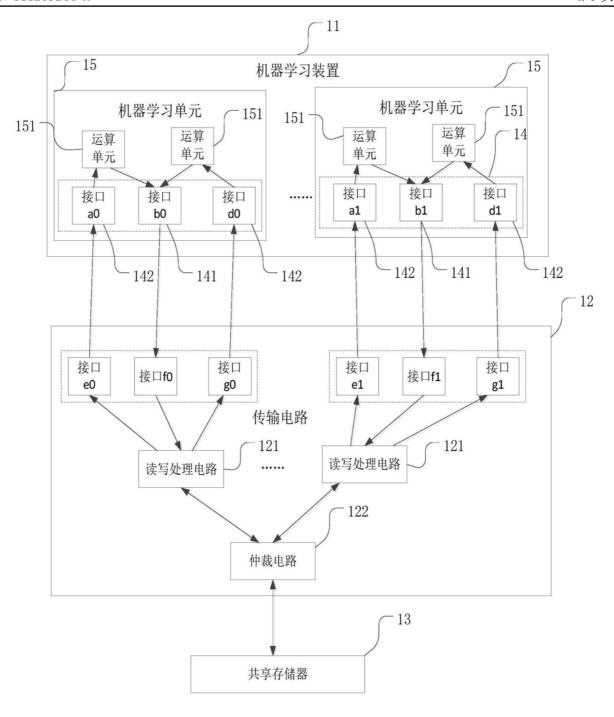
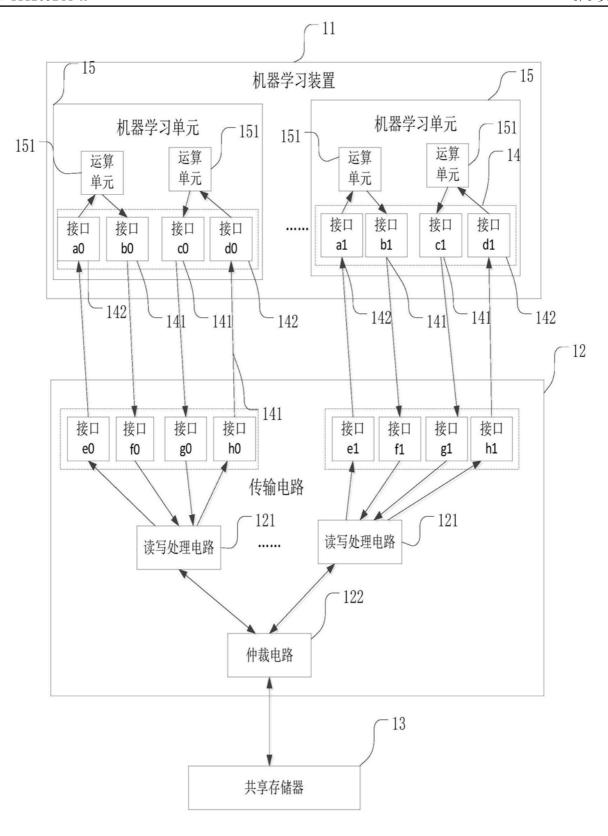
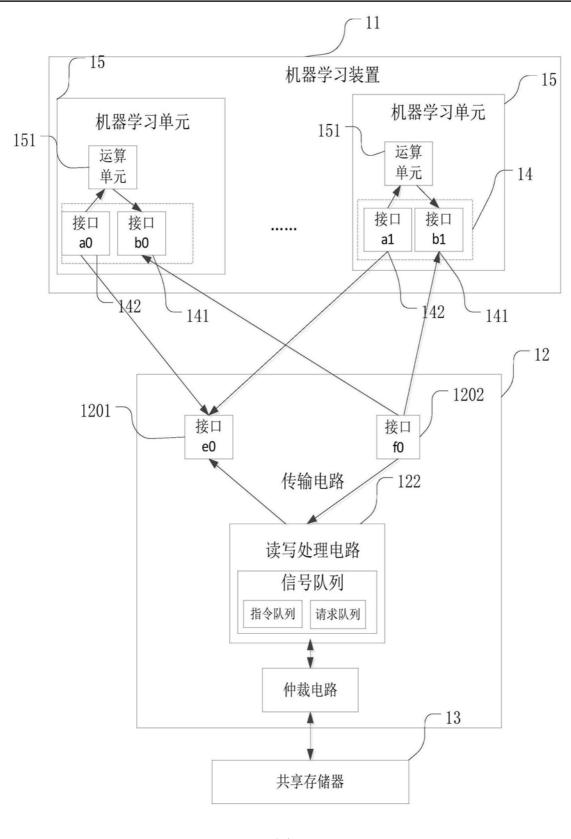


图4





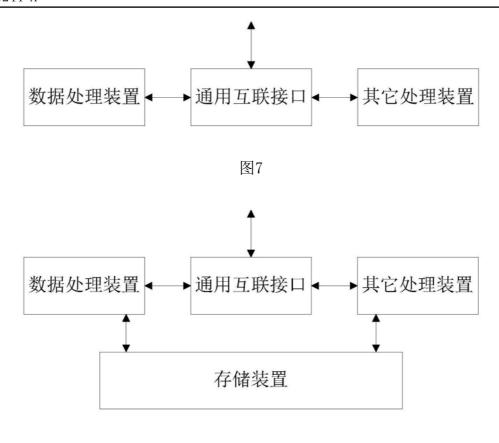


图8

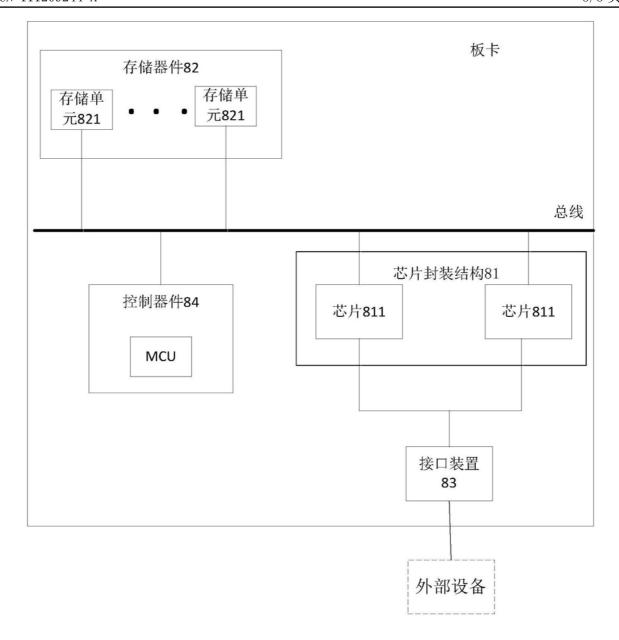


图9