(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112765542 B
(45)授权公告日 2024.11.12

(21)申请号 201911061951.9

(22)申请日 2019.11.01

(65)同一申请的已公布的文献号
申请公布号 CN 112765542 A

(43)申请公布日 2021.05.07

(73)专利权人 中科寒武纪科技股份有限公司
地址 100190 北京市海淀区科学院南路6号科研综合楼644室

(72)发明人 请求不公布姓名 请求不公布姓名
请求不公布姓名 请求不公布姓名
请求不公布姓名 请求不公布姓名
请求不公布姓名 请求不公布姓名
请求不公布姓名 请求不公布姓名

(74)专利代理机构 北京同立钧成知识产权代理有限公司 11205
专利代理师 胡艾青 刘芳

(51)Int.Cl.
G06F 17/15(2006.01)
G06F 17/16(2006.01)

(56)对比文件
CN 109325591 A,2019.02.12

审查员 肖云鹏

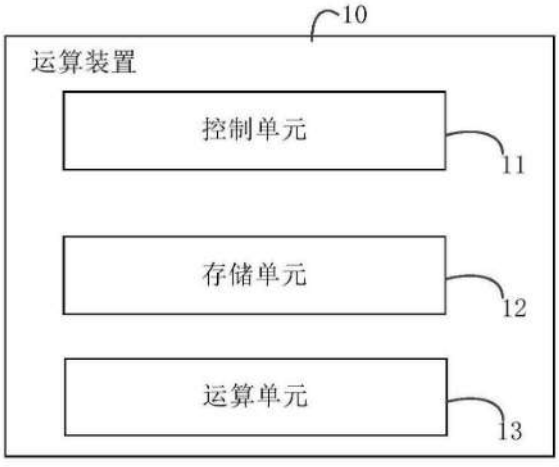
权利要求书4页 说明书15页 附图4页

(54)发明名称

运算装置

(57)摘要

本申请实施例提供一种运算装置中,运算装置包括存储单元、控制单元和运算单元。本申请提供的技术方案可降低卷积运算的资源消耗、提高卷积运算速度、减少运算时间。



CN 112765542 B

1.一种运算装置,所述装置用于进行winograd卷积运算,其特征在于,所述装置包括:控制单元、存储单元、以及运算单元;

所述控制单元,用于发送控制指令,所述控制指令用于指示所述运算单元进行winograd卷积运算;

所述存储单元,用于存储用于winograd卷积运算的数据;

所述运算单元,用于响应所述控制指令,从所述存储单元中提取数据进行winograd卷积运算,其中,所述运算单元将所述winograd卷积运算中对所述数据的变换运算拆解为求和运算,并根据所述求和运算完成所述数据的winograd变换;

所述运算单元,具体用于将所述数据拆解为多个子张量;对所述多个子张量进行变换运算并求和,根据求和运算的结果得到所述数据的winograd变换结果;

所述运算单元,具体用于从所述数据解析得到多个子张量,其中,所述数据为所述多个子张量之和,所述多个子张量的个数与所述数据中非0元素的个数相同,每个所述子张量中有单个非0元素,且在所述子张量中的非0元素与在所述数据中对应位置的非0元素相同。

2.根据权利要求1所述的运算装置,其特征在于,

所述运算单元,具体用于获取各子张量对应的元子张量的winograd变换结果,其中,所述元子张量是将所述子张量的非0元素置为1的张量;将所述子张量中非0的元素值作为系数乘以对应的元子张量的winograd变换结果,得到所述子张量的winograd变换结果;将多个子张量的winograd变换结果相加得到所述数据的winograd变换结果。

3.根据权利要求2所述的运算装置,其特征在于,所述运算单元,具体用于对于每一个所述子张量,将所述子张量对应的元子张量左边乘以左乘矩阵、右边乘以右乘矩阵,得到所述元子张量的winograd变换结果,其中,所述左乘矩阵和所述右乘矩阵都是由所述子张量的规模以及winograd变换类型确定的,其中所述winograd变换类型包括正变换的winograd变换类型和逆变换的winograd变换类型。

4.根据权利要求1至3任一所述的运算装置,其特征在于,

所述数据包括特征数据、权值数据中的至少一种;

所述winograd变换包括正变换和/或逆变换。

5.根据权利要求4所述的运算装置,其特征在于,

所述控制指令包括第一指令和第二指令,其中,所述第一指令包括正变换指令,所述第二指令包括对位乘指令和逆变换指令;

所述运算单元,用于响应所述第一指令,从所述存储单元中提取所述特征数据,对所述特征数据进行正变换,其中,所述运算单元将对所述特征数据的正变换拆解为求和运算,并根据所述求和运算完成所述特征数据的正变换,得到特征变换结果;

所述运算单元,还用于响应所述第二指令获取经过正变换的权值变换结果,对所述权值变换结果和特征变换结果进行对位乘,得到乘法运算结果;对所述乘法运算结果进行逆变换,其中,所述运算单元将对所述乘法运算结果的逆变换拆解为求和运算,并根据所述求和运算完成所述乘法运算结果的逆变换,得到运算结果。

6.根据权利要求5所述的运算装置,其特征在于,

所述存储单元,具体用于接收权值变换结果并存储;

所述运算单元,具体用于响应所述第二指令,从所述存储单元中提取所述权值变换结

果。

7.根据权利要求5所述的运算装置,其特征在于,

所述存储单元,具体用于存储权值数据;

所述运算单元,具体用于从所述存储单元中提取所述权值数据,对所述权值数据进行正变换,其中,所述运算单元将对所述权值数据的正变换拆解为求和运算,并根据求和运算完成所述权值数据的正变换,获得权值变换结果。

8.根据权利要求5所述的运算装置,其特征在于,所述运算单元包括:

第一运算单元,用于响应所述第一指令,从所述存储单元提取特征数据,对所述特征数据进行正变换,其中,所述第一运算单元将对所述特征数据的正变换拆解为求和运算,并根据所述求和运算完成所述特征数据的正变换,得到特征变换结果;

第二运算单元,用于响应所述第二指令获取权值变换结果,对所述权值变换结果和特征变换结果进行对位乘,得到乘法运算结果;对所述乘法运算结果进行逆变换,其中,所述第二运算单元将所述逆变换中对所述乘法运算结果的逆变换运算拆解为求和运算,并根据所述求和运算完成所述乘法运算结果的逆变换,得到运算结果。

9.根据权利要求8所述的运算装置,其特征在于,所述第二运算单元包括:

乘法单元,用于响应所述第二指令,获取权值变换结果,对所述权值变换结果和特征变换结果进行对位乘,得到乘法运算结果;

逆变换单元,用于对所述乘法运算结果进行逆变换,其中,所述逆变换单元将所述逆变换中对所述乘法运算结果的变换运算拆解为求和运算,并根据所述求和运算完成所述乘法运算结果的逆变换,得到所述运算结果。

10.根据权利要求5所述的运算装置,其特征在于,所述运算单元包括:

加法运算单元,用于响应所述第一指令,从所述存储单元获取特征数据,对所述特征数据进行正变换,其中,所述加法运算单元将对所述特征数据的正变换拆解为求和运算,并根据所述求和运算完成所述特征数据的正变换,得到特征变换结果;

乘法运算单元,用于响应所述第二指令获取权值变换结果,对所述权值变换结果和特征变换结果进行对位乘,得到乘法运算结果;

所述加法运算单元,还用于响应所述第二指令,对所述乘法运算结果进行逆变换,其中,所述加法运算单元将对所述乘法运算结果的逆变换拆解为求和运算,并根据所述求和运算完成所述乘法运算结果的逆变换,得到运算结果。

11.一种人工智能芯片,其特征在于,所述芯片包括如权利要求1-10中任意一项所述的运算装置。

12.一种电子设备,其特征在于,所述电子设备包括如权利要求11所述的人工智能芯片。

13.一种板卡,其特征在于,所述板卡包括:存储器件、接口装置和控制器件以及如权利要求11所述的人工智能芯片;

其中,所述人工智能芯片与所述存储器件、所述控制器件以及所述接口装置分别连接;

所述存储器件,用于存储数据;

所述接口装置,用于实现所述人工智能芯片与外部设备之间的数据传输;

所述控制器件,用于对所述人工智能芯片的状态进行监控。

14.根据权利要求13所述的板卡,其特征在于,

所述存储器件包括:多组存储单元,每一组所述存储单元与所述人工智能芯片通过总线连接,所述存储单元为:DDR SDRAM;

所述芯片包括:DDR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;

所述接口装置为:标准PCIE接口。

15.一种运算方法,其特征在于,应用于运算装置,所述运算装置包括:控制单元、存储单元、以及运算单元;其中,

所述控制单元发送控制指令,所述控制指令用于指示所述运算单元进行winograd卷积运算,

所述存储单元存储用于winograd卷积运算的数据;

所述运算单元响应所述控制指令,从所述存储单元中提取数据进行winograd卷积运算,其中,所述运算单元将所述winograd卷积运算中对所述数据的变换运算拆解为求和运算,并根据所述求和运算完成所述数据的winograd变换;

所述运算单元将所述数据拆解为多个子张量;对所述多个子张量进行变换运算并求和,根据求和运算的结果得到所述数据的winograd变换结果;

所述运算单元从所述数据解析得到多个子张量,其中,所述数据为所述多个子张量之和,所述多个子张量的个数与所述数据中非0元素的个数相同,每个所述子张量中有单个非0元素,且在所述子张量中的非0元素与在所述数据中对应位置的非0元素相同。

16.根据权利要求15所述的运算方法,其特征在于,

所述运算单元获取各子张量对应的元子张量的winograd变换结果,其中,所述元子张量是将所述子张量的非0元素置为1的张量;将所述子张量中非0的元素值作为系数乘以对应的元子张量的winograd变换结果,得到所述子张量的winograd变换结果;将多个子张量的winograd变换结果相加得到所述数据的winograd变换结果。

17.根据权利要求16所述的运算方法,其特征在于,所述运算单元对于每一个所述子张量,将所述子张量对应的元子张量左边乘以左乘矩阵、右边乘以右乘矩阵,得到所述元子张量的winograd变换结果,其中,所述左乘矩阵和所述右乘矩阵都是由所述子张量的规模以及winograd变换类型确定的,其中所述winograd变换类型包括正变换的winograd变换类型和逆变换的winograd变换类型。

18.根据权利要求15至17任一所述的运算方法,其特征在于,

所述数据包括特征数据、权值数据中的至少一种;

所述winograd变换包括正变换和/或逆变换。

19.根据权利要求18所述的运算方法,其特征在于,所述控制指令包括第一指令和第二指令,其中,所述第一指令包括正变换指令,所述第二指令包括对位乘指令和逆变换指令;

所述运算单元响应所述第一指令,从所述存储单元中提取所述特征数据,对所述特征数据进行winograd卷积运算,其中,所述运算单元将所述winograd卷积运算中对所述特征数据的变换运算拆解为求和运算,并根据所述求和运算完成所述特征数据的正变换,得到特征变换结果;

所述运算单元还响应所述第二指令获取经过正变换的权值变换结果,对所述权值变换结果和特征变换结果进行对位乘,得到乘法运算结果;对所述乘法运算结果进行逆变换,其

中,所述运算单元将对所述乘法运算结果的逆变换拆解为求和运算,并根据所述求和运算完成所述乘法运算结果的逆变换,得到运算结果。

20.根据权利要求19所述的运算方法,其特征在于,

所述存储单元接收权值变换结果并存储;

所述运算单元响应所述第二指令,从所述存储单元中提取所述权值变换结果。

21.根据权利要求19所述的运算方法,其特征在于,

所述存储单元存储权值数据;

所述运算单元从所述存储单元中提取所述权值数据,对所述权值数据进行正变换,其中,所述运算单元将对所述权值数据的正变换拆解为求和运算,并根据求和运算完成所述权值数据的正变换,获得权值变换结果。

22.根据权利要求19所述的运算方法,其特征在于,所述运算单元包括:第一运算单元和第二运算单元;

所述第一运算单元响应所述第一指令,响应所述第一指令,从所述存储单元提取特征数据,对所述特征数据进行正变换,其中,所述第一运算单元将对所述特征数据的正变换拆解为求和运算,并根据所述求和运算完成所述特征数据的正变换,得到特征变换结果;

所述第二运算单元响应所述第二指令获取权值变换结果,对所述权值变换结果和特征变换结果进行对位乘,得到乘法运算结果;对所述乘法运算结果进行逆变换,其中,所述第二运算单元将对所述乘法运算结果的逆变换拆解为求和运算,并根据所述求和运算完成所述乘法运算结果的逆变换,得到运算结果。

23.根据权利要求22所述的运算方法,其特征在于,所述第二运算单元包括:乘法单元和逆变换单元;

所述乘法单元响应所述第二指令,获取经过正变换的权值变换结果,对所述权值变换结果和特征变换结果进行对位乘,得到乘法运算结果;

所述逆变换单元对所述乘法运算结果进行逆变换,其中,所述逆变换单元将对所述乘法运算结果的逆变换拆解为求和运算,并根据所述求和运算完成所述乘法运算结果的逆变换,得到运算结果。

24.根据权利要求19所述的运算方法,其特征在于,所述运算单元包括:加法运算单元和乘法运算单元;

所述加法运算单元响应所述第一指令,从所述存储单元获取特征数据,对所述特征数据进行正变换,其中,所述加法运算单元将对所述特征数据的正变换拆解为求和运算,并根据所述求和运算完成所述特征数据的正变换,得到特征变换结果;

所述乘法运算单元响应所述第二指令获取权值变换结果,对所述权值变换结果和特征变换结果进行对位乘,得到乘法运算结果;

所述加法运算单元,还用于响应所述第二指令,对所述乘法运算结果进行逆变换,其中,所述加法运算单元将对所述乘法运算结果的逆变换拆解为求和运算,并根据所述求和运算完成所述乘法运算结果的逆变换,得到运算结果。

运算装置

技术领域

[0001] 本申请涉及人工智能技术领域,具体涉及一种运算装置。

背景技术

[0002] 随着人工智能技术的发展,用于实现人工智能技术的各种运算装置和包含运算装置的网络,被广泛应用于计算机视觉、自然语言处理等领域。

[0003] 为了适应越来越高的任务要求,包含运算装置的网络规模越来越大,需要做更大量的运算,尤其是卷积运算。而在进行卷积运算时,现有的运算装置功耗高、计算时间较长,影响其在人工智能技术领域中的应用。

发明内容

[0004] 基于此,有必要针对上述技术问题,提供一种能够提高运算速度、减少运算时间、降低功耗的运算装置。

[0005] 本申请实施例第一方面提供了一种运算装置,所述装置用于进行winograd卷积运算,所述装置包括:控制单元、存储单元、以及运算单元;

[0006] 所述控制单元,用于发送控制指令,所述控制指令用于指示所述运算单元进行winograd卷积运算,

[0007] 所述存储单元,用于存储用于winograd卷积运算的数据;

[0008] 所述运算单元,用于响应所述控制指令,从所述存储单元中提取数据进行winograd卷积运算,其中,所述运算单元将所述winograd卷积运算中对所述数据的变换运算拆解为求和运算,并根据所述求和运算完成所述数据的winograd变换。

[0009] 本申请实施例第二方面提供了一种人工智能芯片,所述芯片包括如本申请第一方面中任意一项所述的运算装置。

[0010] 本申请实施例第三方面提供了一种电子设备,所述电子设备包括如本申请第二方面所述的人工智能芯片。

[0011] 本申请实施例第四方面提供了一种板卡,所述板卡包括:存储器件、接口装置和控制器件以及如本申请第二方面所述的人工智能芯片;

[0012] 其中,所述人工智能芯片与所述存储器件、所述控制器件以及所述接口装置分别连接;

[0013] 所述存储器件,用于存储数据;

[0014] 所述接口装置,用于实现所述人工智能芯片与外部设备之间的数据传输;

[0015] 所述控制器件,用于对所述人工智能芯片的状态进行监控。

[0016] 本申请实施例第五方面提供了一种运算方法,应用于运算装置,所述运算装置包括:控制单元、存储单元、以及运算单元;其中,

[0017] 所述控制单元发送控制指令,所述控制指令用于指示所述运算单元进行winograd卷积运算,

[0018] 所述存储单元存储用于winograd卷积运算的数据；

[0019] 所述运算单元响应所述控制指令，从所述存储单元中提取数据进行winograd卷积运算，其中，所述运算单元将所述winograd卷积运算中对所述数据的变换运算拆解为求和运算，并根据所述求和运算完成所述数据的winograd变换。

[0020] 以上，本申请实施例的方案中，控制单元发送控制指令，控制指令用于指示运算单元进行winograd卷积运算，存储单元存储用于winograd卷积运算的数据；运算单元响应控制指令，从存储单元中提取数据进行winograd卷积运算，其中，运算单元将winograd卷积运算中对数据的变换运算拆解为求和运算，并根据求和运算完成数据的winograd变换，通过以加法运算替代变换运算中的大量乘法运算，加速了winograd卷积运算的速度，也节约了运算资源，本申请提供的方案可降低卷积运算的资源消耗、提高卷积运算速度、减少运算时间。

附图说明

[0021] 为了更清楚地说明本申请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一简单地介绍。

[0022] 图1为一个实施例中运算装置的结构框图；

[0023] 图2为另一个实施例中运算装置的结构框图；

[0024] 图3为再一个实施例中运算装置的结构框图；

[0025] 图4为又一个实施例中运算装置的结构框图；

[0026] 图5为一个实施例中运算方法的流程示意图；

[0027] 图6为一个实施例中板卡的结构框图。

[0028] 其中，图1至图6中的标识如下：

[0029] 10：运算装置；

[0030] 11：控制单元；

[0031] 12：存储单元；

[0032] 13：运算单元；

[0033] 131：第一运算单元；

[0034] 132：第二运算单元；

[0035] 1321：乘法单元；

[0036] 1322：逆变换单元；

[0037] 141：加法运算单元；

[0038] 142：乘法运算单元；

[0039] 389：芯片；

[0040] 390：存储器件；

[0041] 391：接口装置；

[0042] 392：控制器件。

具体实施方式

[0043] 下面将结合本披露实施例中的附图，对本披露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披露一部分实施例，而不是全部的实施例。基于本披露中的实施例，本领域技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本披露保护的范围。

[0044] 应当理解，本披露的权利要求、说明书及附图中的术语“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。本披露的说明书和权利要求书中使用的术语“包括”和“包含”指示所描述特征、整体、步骤、操作、元素和/或组件的存在，但并不排除一个或多个其它特征、整体、步骤、操作、元素、组件和/或其集合的存在或添加。

[0045] 还应当理解，在此本披露说明书中所使用的术语仅仅是出于描述特定实施例的目的，而并不意在限定本披露。如在本披露说明书和权利要求书中所使用的那样，除非上下文清楚地指明其它情况，否则单数形式的“一”、“一个”及“该”意在包括复数形式。还应当进一步理解，在本披露说明书和权利要求书中使用的术语“和/或”是指相关联列出的项中的一个或多个的任何组合以及所有可能组合，并且包括这些组合。

[0046] 为了清楚理解本申请的技术方案，下面对现有技术和本申请实施例中涉及的技术术语进行解释：

[0047] 卷积运算：卷积运算是指从图像的左上角开始，开一个与模板同样大小的活动窗口，活动窗口对应一个窗口图像，该窗口图像为卷积核，窗口图像与图像中的像素对应起来相乘再相加，并用计算结果作为卷积运算后新图像的第一个像素值。然后，活动窗口向右移动一列，并作同样的运算。以此类推，从左到右、从上到下，即可得到一幅新图像。

[0048] winograd卷积：Winograd卷积是一种基于多项式插值算法的卷积加速实现方式。它通过对卷积操作的两个输入：第一目标矩阵和第二目标矩阵分别进行winograd卷积正变换，再将正变换后的第一目标矩阵和第二目标矩阵进行对位乘法，最后对对位乘法结果再次进行winograd卷积逆变换，得到与原卷积操作等价的卷积结果。

[0049] 现有的人工智能技术，通常是基于处理器的卷积运算来实现特征的提取，例如神经网络中对特征数据的运算处理。神经网络中的卷积层对输入的特征数据以预设的卷积核进行卷积处理，并输出运算结果。其中，卷积层具体可以包含有相邻设置的多层卷积层，且每一层卷积层得到的运算结果，为对其上一层卷积层输入的特征数据。

[0050] 针对每一卷积层中的运算，现有的卷积运算中，以包含权值数据的卷积核为窗口，在输入的特征数据的矩阵上“滑动”，对每个滑动位置上的局部矩阵执行矩阵乘，并根据矩阵乘的结果得到最后的运算结果。可见，现有的卷积运算中，需要大量运用到矩阵乘，而矩阵乘需要对特征数据中每行矩阵元素与卷积核的每列矩阵元素，进行对位相乘后累加的处理，其计算量庞大、处理效率低，运算装置能耗也高。

[0051] 为了解决上述问题，本申请实施例提供了一种运算装置，用于进行winograd卷积运算。其中，还进一步将winograd卷积运算中对数据的变换运算拆解为求和运算，并根据该求和运算完成所述数据的winograd变换实现对运算过程的加速。参见图1，为一个实施例中运算装置的结构框图。如图1所示的运算装置10，包括：控制单元11、存储单元12、以及运算单元13。其中，控制单元11通过发出指令而对存储单元12和运算单元13进行控制，最终得到运算结果。在本实施例中，在对特征数据进行正变换的过程中，特征数据就是该过程中的目标矩阵；在对权值数据进行正变换的过程中，权值数据就是该过程中的目标矩阵；在对乘法运算结果进行逆变换的过程中，乘法运算结果就是该过程中的目标矩阵。

[0052] 继续参见图1,控制单元11,用于发送控制指令,所述控制指令用于指示所述运算单元进行winograd卷积运算。例如,控制指令可以包括第一指令和第二指令,其中,所述第一指令包括正变换指令,所述第二指令包括对位乘指令和逆变换指令。控制单元11,用于发送第一指令和第二指令,以控制所述运算单元从存储单元中提取数据,进行相应的winograd变换。

[0053] 存储单元12,用于存储用于winograd卷积运算的数据。该数据例如包括特征数据、权值数据中的至少一种。所述winograd变换包括正变换和/或逆变换。运算单元13,用于响应所述控制指令,从所述存储单元中提取数据进行winograd卷积运算,其中,所述运算单元将所述winograd卷积运算中对所述数据的变换运算拆解为求和运算,并根据所述求和运算完成所述数据的winograd变换。在本申请的各种实施例中,winograd卷积运算可以理解为是采用下式进行计算:

[0054] $S=A^T((GgG^T)\odot(B^TdB))A$

[0055] 其中,S表示卷积矩阵,即使用特征数据与权值数据进行卷积运算得到的结果矩阵;d表示特征数据;g表示权值数据;B表示将特征数据实现正变换的特征变换矩阵;B^T表示B的转置;G表示将权值数据实现正变换的权值变换矩阵;G^T表示G的转置;A表示将对位乘后的乘法运算结果实现逆变换的变换矩阵;A^T表示A的转置。

[0056] 由上式可知,在winograd卷积运算中需要进行多次winograd变换(正变换或逆变换)而这些winograd变换涉及大量乘法运算。本申请通过将winograd卷积运算中对数据(例如特征数据)的变换运算拆解为求和运算,并根据求和运算完成所述数据的winograd变换,以加法运算替代变换运算中的大量乘法运算,加速了winograd卷积运算的速度,也节约了运算资源,本申请提供的方案可降低卷积运算的资源消耗、提高卷积运算速度、减少运算时间。

[0057] 在一些实施例中,所述运算单元13,具体用于将所述数据拆解为多个子张量;对所述多个子张量进行变换运算并求和,根据求和运算的结果得到所述数据的winograd变换结果。

[0058] 将数据拆解为多个子张量的过程,可以理解为:运算单元,具体用于从所述数据解析得到多个子张量,其中,所述数据为所述多个子张量之和,所述多个子张量的个数与所述数据中非0元素的个数相同,每个所述子张量中有单个非0元素,且在所述子张量中的非0元素与在所述数据中对应位置的非0元素相同。

[0059] 以下面X×Y规模的特征数据d为例,对运算单元从数据解析得到多个子张量的过程。

[0060] $$d=\begin{bmatrix} d_{00} & d_{01} & \cdots\cdots & d_{0Y} \\ d_{10} & d_{11} & \cdots\cdots & d_{1Y} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ d_{X0} & d_{X1} & \cdots\cdots & d_{XY} \end{bmatrix}$$

[0061] 将上述特征数据d拆解为多个子张量:

[0062] $$\begin{bmatrix} d_{00} & d_{01} & \cdots & d_{0Y} \\ d_{10} & d_{11} & \cdots & d_{1Y} \\ \cdots & \cdots & \cdots & \cdots \\ d_{X0} & d_{X1} & \cdots & d_{XY} \end{bmatrix} = \begin{bmatrix} d_{00} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} + \cdots + \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & 0 \\ 0 & 0 & 0 & d_{XY} \end{bmatrix}$$

[0063] 假设特征数据d中所有元素都是非0元素，由此，得到X×Y个子张量。在其他实施例中，假如特征数据d中仅有3个元素是非0元素，则只能得到3个子张量，在此不做具体举例。

[0064] $$\begin{bmatrix} d_{00} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}、\begin{bmatrix} 0 & d_{01} & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}、\cdots、\begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & 0 \\ 0 & 0 & 0 & d_{XY} \end{bmatrix}$$

[0065] 在上述X×Y个子张量中，以第一个子张量为例，其具有单个非0元素d_{00}，其余元素均为0，且该子张量的这个非0元素，与特征数据中对应位置的元素相同，均为d_{00}。

[0066] 在拆解得到多个子张量后，运算单元对所述多个子张量进行变换运算并求和，根据求和运算的结果得到所述数据的winograd变换结果的过程，具体可以是：运算单元获取各子张量对应的元子张量的winograd变换结果，其中，所述元子张量是将所述子张量的非0元素置为1的张量；将所述子张量中非0的元素值作为系数乘以对应的元子张量的winograd变换结果，得到所述子张量的winograd变换结果；将多个子张量的winograd变换结果相加得到所述数据的winograd变换结果。

[0067] 继续以上述特征数据d为例进行举例，以第一个子张量为例，将其中的非0元素d_{00}置为1后可以得到如下的元子张量：

[0068] $$\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

[0069] 对第一个子张量的元子张量可获取到如下的winograd变换结果：

[0070] $$B_{X\times X}^{T}\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}B_{Y\times Y}$$

[0071] 其中，运算单元，具体用于对于每一个所述子张量，将所述子张量对应的元子张量左边乘以左乘矩阵（例如$B_{X\times X}^{T}$）、右边乘以右乘矩阵（例如$B_{Y\times Y}$），得到所述元子张量的winograd变换结果。其中，所述左乘矩阵和所述右乘矩阵都是由所述子张量的规模以及winograd变换类型确定的，其中所述winograd变换类型包括正变换的winograd变换类型和逆变换的winograd变换类型。

[0072] 在具体的实现方式中，用于将数据实现正变换或逆变换的变换矩阵（例如B、G、A）都是由数据的规模而确定的，不同规模的数据对应的变换矩阵都是预设的已知矩阵。因此，对上述第一个子张量的元子张量的winograd变换结果，可以理解为一常数矩阵。然后，将X

×Y个子张量对应的X×Y个winograd变换结果相加，得到如下特征数据的winograd变换结果。

$$(B^T dB)=d_{00}B^T_{X\times X}\begin{bmatrix}1 & 0 & \cdots\cdots & 0\\ 0 & 0 & \cdots\cdots & 0\\ \cdots\cdots & \cdots\cdots & \cdots\cdots & 0\\ 0 & 0 & 0 & 0\end{bmatrix}B_{Y\times Y}+\cdots+$$

[0073]

$$d_{XY}B^T_{X\times X}\begin{bmatrix}0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 1\end{bmatrix}B_{Y\times Y}$$

[0074] 所述winograd变换包括正变换和/或逆变换，上述实施例中是以特征数据的winograd变换，对变换运算拆解为求和运算进行举例，但上述拆解方式也可以用于权值数据正变换运算（(GgG^T)）、以及$A(GgG^T)$与(B^TdB)对位乘的乘法运算结果逆变换运算中，在此不做赘述。

[0075] 在一些实施例中，所述控制指令包括第一指令和第二指令，其中，所述第一指令包括正变换指令，所述第二指令包括对位乘指令和逆变换指令。控制单元11，用于发出第一指令和第二指令。在一些实施例中，控制单元11发出的第一指令和第二指令可以是预先从存储单元12中提取的，也可以是预先由外部写入并存储在控制单元11内的。例如，第一指令和第二指令都包括操作码和操作地址。第一指令包括与正变换指令相对应的正变换操作码、操作地址。第二指令包括与对位乘指令相对应的对位乘操作码、操作地址，以及与逆变换指令相对应的逆变换操作码、操作地址。每条指令可以包括一个操作码以及一个或多个操作地址。其中，操作地址，具体可以是寄存器地址。

[0076] 图1所示的存储单元12存储数据。存储单元12存储的数据例如可以是运算装置10在winograd卷积运算中需要用到的数据。在上述第一指令和第二指令是预先从存储单元12中提取的实施例中，存储单元12存储的数据中包含了第一指令和第二指令。存储单元12的结构例如可以包括寄存器、缓存和数据输入/输出单元。

[0077] 继续参见图1，运算单元13，用于响应所述第一指令，从所述存储单元12中提取所述特征数据，对所述特征数据进行正变换，其中，所述运算单元13将对所述特征数据的正变换拆解为求和运算，并根据求和运算完成所述特征数据的正变换，得到特征变换结果。具体地，例如，运算单元13获取到控制单元11发来的第一指令时，运算单元13从第一指令解析得到对特征数据进行正变换的指令。运算单元13从存储单元12读取特征数据，对该特征数据进行正变换，得到特征变换结果。其中，运算单元13还可以根据特征数据的规模，从存储单元12获取到与该特征数据相对应的特征变换矩阵。

[0078] 继续参见图1，运算单元13，还用于响应所述第二指令获取经过正变换的权值变换结果，对所述权值变换结果和特征变换结果进行对位乘，得到乘法运算结果；对所述乘法运算结果进行逆变换，其中，所述运算单元将对所述乘法运算结果的逆变换拆解为求和运算，并根据所述求和运算完成所述乘法运算结果的逆变换，得到运算结果。

[0079] 具体地，例如，运算单元13获取到控制单元11发来的第二指令时，可以是获取到对位乘指令和逆变换指令。其中，运算单元13根据对位乘指令获取权值变换结果和特征变换结果，并对两者进行对位乘。其中，可以是在得到特征变换结果时，从存储单元12获取预先

存储的权值变换结果，进行对位乘；也可以是同时计算得到权值变换结果和特征变换结果，然后对两者进行对位乘。对位乘是两个矩阵行列相同位置元素一一对应相乘得到的乘法运算结果，并不改变矩阵规模。在对位乘之后，运算单元13根据逆变换指令，获取与乘法运算结果相对应的逆变换矩阵（例如A），并以该逆变换矩阵对乘法运算结果进行逆变换，得到运算结果。示例性的，如果特征数据是待处理图像的特征数据，那么运算单元13得到的运算结果可以理解为对待处理图像的特征提取结果。上述对乘法运算结果进行逆变换的过程中，都可以将对所述乘法运算结果的逆变换拆解为求和运算，并根据求和运算完成所述乘法运算结果的逆变换，得到运算结果，拆解方式与前述实施例中对特征数据正变换的拆解方式相同，在此不做赘述。

[0080] 以上运算装置，通过对特征数据的正变换得到特征变换结果，以及对权值变换结果和特征变换结果进行对位乘和逆变换，将对所述乘法运算结果的逆变换拆解为求和运算，以加法运算替代了现有卷积运算过程中的大量乘法运算，通过减少乘法运算加速运算速度、减少了资源消耗。

[0081] 在上述实施例中，权值变换结果可以是与特征变换结果同时计算得到的，也可以是先于特征变换结果得到而预先存储的。

[0082] 在一些实施例中，存储单元12，具体用于存储权值数据。图1所示的运算单元13，具体用于从存储单元12提取所述权值数据，对所述权值数据进行正变换，其中，所述运算单元将对所述权值数据的正变换拆解为求和运算，并根据求和运算完成所述权值数据的正变换，获得权值变换结果。

[0083] 在另一些实施例中，存储单元12，具体用于接收权值变换结果并存储。运算单元13，具体用于响应所述第二指令，从所述存储单元12中获取所述权值变换结果。示例性的，本实施例在对权值数据进行预先存储时或者是接收到第一指令之前，运算单元13可以预先对所述权值数据进行正变换，获得权值变换结果，并将权值变换结果存入存储单元12。然后，运算单元13响应第一指令对特征数据进行正变换，得到特征变换结果。由此，可以直接提取权值变换结果，减少winograd卷积运算的运算时间。最后，运算单元13响应第二指令，从存储单元12中提取对所述权值变换结果和特征变换结果进行对位乘，得到乘法运算结果；对所述乘法运算结果进行逆变换，其中，运算单元13将对所述乘法运算结果的逆变换拆解为求和运算，并根据所述求和运算完成所述乘法运算结果的逆变换，得到运算结果。可选地，运算单元13对特征数据进行正变换得到特征变换结果的过程，和运算单元13对权值变换结果和特征变换结果进行对位乘的过程，可以同步执行，以提高运算速度和效率。

[0084] 又可选地，运算单元13响应第一指令，从存储单元12获取权值数据和特征数据，先分别对权值数据和特征数据进行正变换，得到权值变换结果和特征变换结果。然后，运算单元13响应第二指令对权值变换结果和特征变换结果进行对位乘和逆变换。

[0085] 对于上述实施例中的运算单元13，具体结构可以是多种，例如，一种运算单元13可以包括第一运算单元和第二运算单元；另一种运算单元13可以包括加法运算单元和乘法运算单元，下面结合附图对这两种可能的结构进行举例说明。

[0086] 参见图2，为另一个实施例中运算装置的结构框图。如图2所示的运算装置10，运算单元13可以包括第一运算单元131和第二运算单元132。

[0087] 其中，第一运算单元131，用于响应所述第一指令，从所述存储单元提取特征数据，

对所述特征数据进行正变换，其中，第一运算单元131将对所述特征数据的正变换拆解为求和运算，并根据所述求和运算完成所述特征数据的正变换，得到特征变换结果。在一些实施例中，第一运算单元131在对所述特征数据进行正变换时，还可以同时从所述存储单元12获取权值数据，对权值数据进行正变换，得到权值变换结果。然后，将得到的特征变换结果和权值变换结果都发送至第二运算单元132。在另一些实施例中，由于第一运算单元131与第二运算单元132之间传输带宽有限，为了降低带宽占用量，第一运算单元131在响应第一指令对所述特征数据进行正变换之前，例如在接收到第一指令之前，可以预先对权值数据进行正变换，得到权值变换结果，并将权值变换结果存储在存储单元12中。在第一运算单元131将特征变换结果发给第二运算单元132时，第二运算单元132可以直接从存储单元12中获取到权值变换结果。由此，第一运算单元131与第二运算单元132之间不需要传输权值变换结果，降低了对第一运算单元131与第二运算单元132之间传输带宽的要求、提高了传输速度。

[0088] 第二运算单元132，用于响应所述第二指令获取权值变换结果，对所述权值变换结果和特征变换结果进行对位乘，得到乘法运算结果；对所述乘法运算结果进行逆变换，其中，第二运算单元132将所述逆变换中对所述乘法运算结果的逆变换运算拆解为求和运算，并根据求和运算完成所述乘法运算结果的逆变换，得到运算结果。

[0089] 参见图3，为再一个实施例中运算装置的结构框图。在图3所示的运算装置10中，第二运算单元132具体可以包括乘法单元1321和逆变换单元1322。

[0090] 其中，乘法单元1321，用于响应所述第二指令，获取权值变换结果，对所述权值变换结果和特征变换结果进行对位乘，得到乘法运算结果。图3所示乘法单元1321，可以是从存储单元12获取预先存储的权值变换结果，或者是从第一运算单元131获取其计算得到的权值变换结果，在此不做限定。

[0091] 逆变换单元1322，用于对所述乘法运算结果进行逆变换，其中，逆变换单元1322将所述逆变换中对所述乘法运算结果的逆变换拆解为求和运算，并根据所述求和运算完成所述乘法运算结果的逆变换，得到所述运算结果。逆变换单元1322具体可以从存储单元12获取用于对乘法运算结果进行逆变换逆变换矩阵(例如A)，并以逆变换矩阵对对所述乘法运算结果进行逆变换，得到运算结果。

[0092] 参见图4，为又一个实施例中运算装置的结构框图。在图4所示的运算装置10中，运算单元13包括加法运算单元141和乘法运算单元142。在本实施例中，将运算过程中能够以加法完成的过程用加法运算单元141执行，而将对位乘用专门的乘法运算单元142执行。

[0093] 加法运算单元141，用于响应所述第一指令，从所述存储单元12获取特征数据，对所述特征数据进行正变换，其中，加法运算单元141将对所述特征数据的正变换拆解为求和运算，并根据所述求和运算完成所述特征数据的正变换，得到特征变换结果其中，加法运算单元141可以在接收到第一指令之前，预先对权值数据进行正变换，得到权值变换结果，并将权值变换结果存入存储单元12。由此，加法运算单元141和乘法运算单元142之间不需要传输权值变换结果，降低了对传输带宽的要求、提高了传输速度。或者，加法运算单元141可以响应第一指令，对所述特征数据进行正变换的同时，对权值数据进行正变换，得到权值变换结果后，与特征变换结果一起传输给乘法运算单元142。权值数据可以是存储在存储单元12中的数据。

[0094] 乘法运算单元142,用于响应所述第二指令获取权值变换结果,对所述权值变换结果和特征变换结果进行对位乘,获得乘法运算结果。乘法运算单元142对权值变换结果和特征变换结果中行列相同的元素一一对应做乘法,获得乘法运算结果。例如对4×4的正变换后权值矩阵和特征数据变换结果,一共需要执行16次乘法,获得4×4的乘法运算结果。

[0095] 加法运算单元141,还用于响应所述第二指令,对所述乘法运算结果进行逆变换,其中,所述加法运算单元141将对所述乘法运算结果的逆变换拆解为求和运算,并根据所述求和运算完成所述乘法运算结果的逆变换,得到运算结果。加法运算单元141从乘法运算单元142获取乘法运算结果,并且可以从存储单元12获取逆变换矩阵,以逆变换矩阵对乘法运算结果进行逆变换,得到运算结果。

[0096] 参见图5,为一个实施例中运算方法的流程示意图。图5所示的运算方法应用于上述实施例中的运算装置,所述运算装置包括:控制单元、存储单元、以及运算单元。其中,图5所示的运算方法包括:

[0097] S101,控制单元发送控制指令,所述控制指令用于指示所述运算单元进行winograd卷积运算。

[0098] S102,存储单元存储用于winograd卷积运算的数据。

[0099] S103,运算单元响应所述控制指令,从所述存储单元中提取数据进行winograd卷积运算,其中,所述运算单元将所述winograd卷积运算中对所述数据的变换运算拆解为求和运算,并根据所述求和运算完成所述数据的winograd变换。

[0100] 本申请提供的运算方法中,通过控制单元发送控制指令,控制指令用于指示运算单元进行winograd卷积运算,存储单元存储用于winograd卷积运算的数据;运算单元响应控制指令,从存储单元中提取数据进行winograd卷积运算,其中,运算单元将winograd卷积运算中对数据的变换运算拆解为求和运算,并根据求和运算完成数据的winograd变换,通过以加法运算替代变换运算中的大量乘法运算,加速了winograd卷积运算的速度,也节约了运算资源,本申请提供的方案可降低卷积运算的资源消耗、提高卷积运算速度、减少运算时间。

[0101] 在一些实施例中,所述运算单元将所述数据拆解为多个子张量;对所述多个子张量进行变换运算并求和,根据求和运算的结果得到所述数据的winograd变换结果。

[0102] 在一些实施例中,所述运算单元从所述数据解析得到多个子张量,其中,所述数据为所述多个子张量之和,所述多个子张量的个数与所述数据中非0元素的个数相同,每个所述子张量中有单个非0元素,且在所述子张量中的非0元素与在所述数据中对应位置的非0元素相同。

[0103] 在一些实施例中,所述运算单元获取各子张量对应的元子张量的winograd变换结果,其中,所述元子张量是将所述子张量的非0元素置为1的张量;将所述子张量中非0的元素值作为系数乘以对应的元子张量的winograd变换结果,得到所述子张量的winograd变换结果;将多个子张量的winograd变换结果相加得到所述数据的winograd变换结果。

[0104] 在一些实施例中,所述运算单元对于每一个所述子张量,将所述子张量对应的元子张量左边乘以左乘矩阵、右边乘以右乘矩阵,得到所述元子张量的winograd变换结果,其中,所述左乘矩阵和所述右乘矩阵都是由所述子张量的规模以及winograd变换类型确定的,其中所述winograd变换类型包括正变换的winograd变换类型和逆变换的winograd变换

类型。

[0105] 在一些实施例中,所述数据包括特征数据、权值数据中的至少一种;

[0106] 所述winograd变换包括正变换和/或逆变换。

[0107] 在一些实施例中,所述控制指令包括第一指令和第二指令,其中,所述第一指令包括正变换指令,所述第二指令包括对位乘指令和逆变换指令;

[0108] 所述运算单元响应所述第一指令,从所述存储单元中提取所述特征数据,对所述特征数据进行winograd卷积运算,其中,所述运算单元将所述winograd卷积运算中对所述特征数据的变换运算拆解为求和运算,并根据所述求和运算完成所述特征数据的正变换,得到特征变换结果;

[0109] 所述运算单元还响应所述第二指令获取经过正变换的权值变换结果,对所述权值变换结果和特征变换结果进行对位乘,得到乘法运算结果;对所述乘法运算结果进行逆变换,其中,所述运算单元将对所述乘法运算结果的逆变换拆解为求和运算,并根据所述求和运算完成所述乘法运算结果的逆变换,得到运算结果。

[0110] 在一些实施例中,所述存储单元接收权值变换结果并存储;

[0111] 所述运算单元响应所述第二指令,从所述存储单元中提取所述权值变换结果。

[0112] 在一些实施例中,所述存储单元存储权值数据;

[0113] 所述运算单元从所述存储单元中提取所述权值数据,对所述权值数据进行正变换,其中,所述运算单元将对所述权值数据的正变换拆解为求和运算,并根据求和运算完成所述权值数据的正变换,获得权值变换结果。

[0114] 在一些实施例中,所述运算单元包括:第一运算单元和第二运算单元;

[0115] 所述第一运算单元响应所述第一指令,响应所述第一指令,从所述存储单元提取特征数据,对所述特征数据进行正变换,其中,所述第一运算单元将对所述特征数据的正变换拆解为求和运算,并根据所述求和运算完成所述特征数据的正变换,得到特征变换结果;

[0116] 所述第二运算单元响应所述第二指令获取权值变换结果,对所述权值变换结果和特征变换结果进行对位乘,得到乘法运算结果;对所述乘法运算结果进行逆变换,其中,所述第二运算单元将对所述乘法运算结果的逆变换拆解为求和运算,并根据所述求和运算完成所述乘法运算结果的逆变换,得到运算结果。

[0117] 在一些实施例中,所述第二运算单元包括:乘法单元和逆变换单元;

[0118] 所述乘法单元响应所述第二指令,获取经过正变换的权值变换结果,对所述权值变换结果和特征变换结果进行对位乘,得到乘法运算结果;

[0119] 所述逆变换单元对所述乘法运算结果进行逆变换,其中,所述逆变换单元将对所述乘法运算结果的逆变换拆解为求和运算,并根据所述求和运算完成所述乘法运算结果的逆变换,得到运算结果。

[0120] 在一些实施例中,所述运算单元包括:加法运算单元和乘法运算单元;

[0121] 所述加法运算单元响应所述第一指令,从所述存储单元获取特征数据,对所述特征数据进行正变换,其中,所述加法运算单元将对所述特征数据的正变换拆解为求和运算,并根据所述求和运算完成所述特征数据的正变换,得到特征变换结果;

[0122] 所述乘法运算单元响应所述第二指令获取权值变换结果,对所述权值变换结果和特征变换结果进行对位乘,得到乘法运算结果;

[0123] 所述加法运算单元，还用于响应所述第二指令，对所述乘法运算结果进行逆变换，其中，所述加法运算单元将对所述乘法运算结果的逆变换拆解为求和运算，并根据所述求和运算完成所述乘法运算结果的逆变换，得到运算结果。

[0124] 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，本公开并不受所描述的动作顺序的限制，因为依据本公开，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公开所必须的。

[0125] 进一步需要说明的是，虽然流程图中的各个步骤按照箭头的指示依次显示，但是这些步骤并不是必然按照箭头指示的顺序依次执行。除非本文中有明确的说明，这些步骤的执行并没有严格的顺序限制，这些步骤可以以其它的顺序执行。而且，流程图中的至少一部分步骤可以包括多个子步骤或者多个阶段，这些子步骤或者阶段并不必然是在同一时刻执行完成，而是可以在不同的时刻执行，这些子步骤或者阶段的执行顺序也不必然是依次进行，而是可以与其它步骤或者其它步骤的子步骤或者阶段的至少一部分轮流或者交替地执行。

[0126] 应该理解，上述的装置实施例仅是示意性的，本公开的装置还可通过其它的方式实现。例如，上述实施例中所述单元/模块的划分，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、模块或组件可以结合，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。

[0127] 另外，若无特别说明，在本公开各个实施例中的各功能单元/模块可以集成在一个单元/模块中，也可以是各个单元/模块单独物理存在，也可以两个或两个以上单元/模块集成在一起。上述集成的单元/模块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。

[0128] 所述集成的单元/模块如果以硬件的形式实现时，该硬件可以是数字电路，模拟电路等等。硬件结构的物理实现包括但不局限于晶体管，忆阻器等等。若无特别说明，所述人工智能处理器可以是任何适当的硬件处理器，比如CPU、GPU、FPGA、DSP和ASIC等等。若无特别说明，所述存储单元可以是任何适当的磁存储介质或者磁光存储介质，比如，阻变式存储器RRAM(Resist ive Random Access Memory)、动态随机存取存储器DRAM(Dynamic Random Access Memory)、静态随机存取存储器SRAM(Static Random-Access Memory)、增强动态随机存取存储器EDRAM(Enhanced Dynamic Random Access Memory)、高带宽内存HBM(High-Bandwidth Memory)、混合存储立方HMC(Hybrid Memory Cube)等等。

[0129] 所述集成的单元/模块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，本公开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)执行本公开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。

[0130] 在一种可能的实现方式中,还公开了一种人工智能芯片,其包括了上述运算装置。

[0131] 在一种可能的实现方式中,还公开了一种板卡,其包括存储器件、接口装置和控制器件以及上述人工智能芯片;其中,所述人工智能芯片与所述存储器件、所述控制器件以及所述接口装置分别连接;所述存储器件,用于存储数据;所述接口装置,用于实现所述人工智能芯片与外部设备之间的数据传输;所述控制器件,用于对所述人工智能芯片的状态进行监控。

[0132] 图6示出根据本公开实施例的板卡的结构框图,参阅图6,上述板卡除了包括上述芯片389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390、接口装置391和控制器件392;

[0133] 所述存储器件390与所述人工智能芯片通过总线连接,用于存储数据。所述存储器件可以包括多组存储单元393。每一组所述存储单元与所述人工智能芯片通过总线连接。可以理解,每一组所述存储单元可以是DDR SDRAM(英文:Double Data Rate SDRAM,双倍速率同步动态随机存储器)。

[0134] 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,所述人工智能芯片内部可以包括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。可以理解,当每一组所述存储单元中采用DDR4-3200颗粒时,数据传输的理论带宽可达到25600MB/s。

[0135] 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。

[0136] 所述接口装置与所述人工智能芯片电连接。所述接口装置用于实现所述人工智能芯片与外部设备(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,所述接口装置可以为标准PCIE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,所述接口装置还可以是其他的接口,本公开并不限制上述其他的接口的具体表现形式,所述接口单元能够实现转接功能即可。另外,所述人工智能芯片的计算结果仍由所述接口装置传送回外部设备(例如服务器)。

[0137] 所述控制器件与所述人工智能芯片电连接。所述控制器件用于对所述人工智能芯片的状态进行监控。具体的,所述人工智能芯片与所述控制器件可以通过SPI接口电连接。所述控制器件可以包括单片机(Micro Controller Unit,MCU)。如所述人工智能芯片可以包括多个处理芯片、多个处理核或多个处理电路,可以带动多个负载。因此,所述人工智能芯片可以处于多负载和轻负载等不同的工作状态。通过所述控制装置可以实现对所述人工智能芯片中多个处理芯片、多个处理和或多个处理电路的工作状态的调控。

[0138] 在一种可能的实现方式中,公开了一种电子设备,其包括了上述人工智能芯片。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。

[0139] 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。

[0140] 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。上述实施例的各技术特征可以进行任意的组合,为使描述简洁,未对上述实施例中的各个技术特征所有可能的组合都进行描述,然而,只要这些技术特征的组合不存在矛盾,都应当认为是本说明书记载的范围。

[0141] 依据以下条款可更好地理解前述内容:

[0142] A1、一种运算方法,应用于运算装置,所述运算装置包括:控制单元、存储单元、以及运算单元;其中,

[0143] 所述控制单元发送控制指令,所述控制指令用于指示所述运算单元进行winograd卷积运算,

[0144] 所述存储单元存储用于winograd卷积运算的数据;

[0145] 所述运算单元响应所述控制指令,从所述存储单元中提取数据进行winograd卷积运算,其中,所述运算单元将所述winograd卷积运算中对所述数据的变换运算拆解为求和运算,并根据所述求和运算完成所述数据的winograd变换。

[0146] A2、根据条款A1所述的方法,所述运算单元将所述数据拆解为多个子张量;对所述多个子张量进行变换运算并求和,根据求和运算的结果得到所述数据的winograd变换结果。

[0147] A3、根据条款A2所述的运算方法,

[0148] 所述运算单元从所述数据解析得到多个子张量,其中,所述数据为所述多个子张量之和,所述多个子张量的个数与所述数据中非0元素的个数相同,每个所述子张量中有单个非0元素,且在所述子张量中的非0元素与在所述数据中对应位置的非0元素相同。

[0149] A4、根据条款A2所述的运算方法,

[0150] 所述运算单元获取各子张量对应的元子张量的winograd变换结果,其中,所述元子张量是将所述子张量的非0元素置为1的张量;将所述子张量中非0的元素值作为系数乘以对应的元子张量的winograd变换结果,得到所述子张量的winograd变换结果;将多个子张量的winograd变换结果相加得到所述数据的winograd变换结果。

[0151] A5、根据条款A4所述的运算方法,所述运算单元对于每一个所述子张量,将所述子张量对应的元子张量左边乘以左乘矩阵、右边乘以右乘矩阵,得到所述元子张量的winograd变换结果,其中,所述左乘矩阵和所述右乘矩阵都是由所述子张量的规模以及winograd变换类型确定的,其中所述winograd变换类型包括正变换的winograd变换类型和逆变换的winograd变换类型。

[0152] A6、根据条款A1至A5任一所述的运算方法,

[0153] 所述数据包括特征数据、权值数据中的至少一种;

[0154] 所述winograd变换包括正变换和/或逆变换。

[0155] A7、根据条款A6所述的运算方法,所述控制指令包括第一指令和第二指令,其中,所述第一指令包括正变换指令,所述第二指令包括对位乘指令和逆变换指令;

[0156] 所述运算单元响应所述第一指令,从所述存储单元中提取所述特征数据,对所述

特征数据进行winograd卷积运算，其中，所述运算单元将所述winograd卷积运算中对所述特征数据的变换运算拆解为求和运算，并根据所述求和运算完成所述特征数据的正变换，得到特征变换结果；

[0157] 所述运算单元还响应所述第二指令获取经过正变换的权值变换结果，对所述权值变换结果和特征变换结果进行对位乘，得到乘法运算结果；对所述乘法运算结果进行逆变换，其中，所述运算单元将对所述乘法运算结果的逆变换拆解为求和运算，并根据所述求和运算完成所述乘法运算结果的逆变换，得到运算结果。

[0158] A8、根据条款A7所述的运算方法，

[0159] 所述存储单元接收权值变换结果并存储；

[0160] 所述运算单元响应所述第二指令，从所述存储单元中提取所述权值变换结果。

[0161] A9、根据条款A7所述的运算方法，

[0162] 所述存储单元存储权值数据；

[0163] 所述运算单元从所述存储单元中提取所述权值数据，对所述权值数据进行正变换，其中，所述运算单元将对所述权值数据的正变换拆解为求和运算，并根据求和运算完成所述权值数据的正变换，获得权值变换结果。

[0164] A10、根据条款A7所述的运算方法，所述运算单元包括：第一运算单元和第二运算单元；

[0165] 所述第一运算单元响应所述第一指令，响应所述第一指令，从所述存储单元提取特征数据，对所述特征数据进行正变换，其中，所述第一运算单元将对所述特征数据的正变换拆解为求和运算，并根据所述求和运算完成所述特征数据的正变换，得到特征变换结果；

[0166] 所述第二运算单元响应所述第二指令获取权值变换结果，对所述权值变换结果和特征变换结果进行对位乘，得到乘法运算结果；对所述乘法运算结果进行逆变换，其中，所述第二运算单元将对所述乘法运算结果的逆变换拆解为求和运算，并根据所述求和运算完成所述乘法运算结果的逆变换，得到运算结果。

[0167] A11、根据条款A10所述的运算方法，所述第二运算单元包括：乘法单元和逆变换单元；

[0168] 所述乘法单元响应所述第二指令，获取经过正变换的权值变换结果，对所述权值变换结果和特征变换结果进行对位乘，得到乘法运算结果；

[0169] 所述逆变换单元对所述乘法运算结果进行逆变换，其中，所述逆变换单元将对所述乘法运算结果的逆变换拆解为求和运算，并根据所述求和运算完成所述乘法运算结果的逆变换，得到运算结果。

[0170] A12、根据条款A7所述的运算方法，所述运算单元包括：加法运算单元和乘法运算单元；

[0171] 所述加法运算单元响应所述第一指令，从所述存储单元获取特征数据，对所述特征数据进行正变换，其中，所述加法运算单元将对所述特征数据的正变换拆解为求和运算，并根据所述求和运算完成所述特征数据的正变换，得到特征变换结果；

[0172] 所述乘法运算单元响应所述第二指令获取权值变换结果，对所述权值变换结果和特征变换结果进行对位乘，得到乘法运算结果；

[0173] 所述加法运算单元，还用于响应所述第二指令，对所述乘法运算结果进行逆变换，

其中，所述加法运算单元将对所述乘法运算结果的逆变换拆解为求和运算，并根据所述求和运算完成所述乘法运算结果的逆变换，得到运算结果。

[0174] 以上对本公开实施例进行了详细介绍，本文中应用了具体个例对本公开的原理及实施方式进行了阐述，以上实施例的说明仅用于帮助理解本公开的方法及其核心思想。同时，本领域技术人员依据本公开的思想，基于本公开的具体实施方式及应用范围上做出的改变或变形之处，都属于本公开保护的范围。综上所述，本说明书内容不应理解为对本公开的限制。

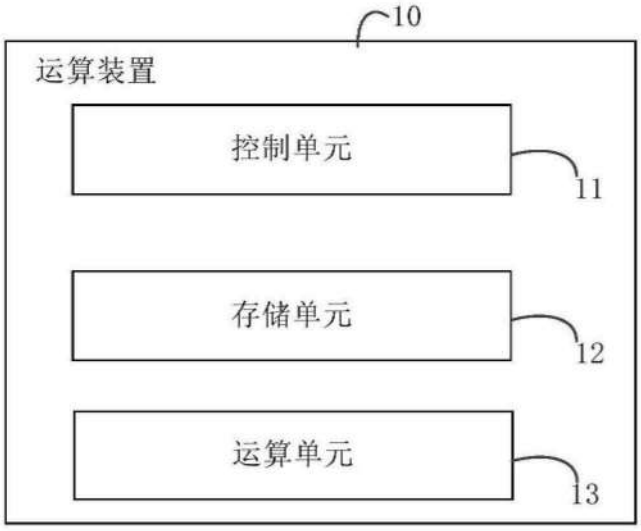

10
运算装置
控制单元
11
存储单元
12
运算单元
13

图1

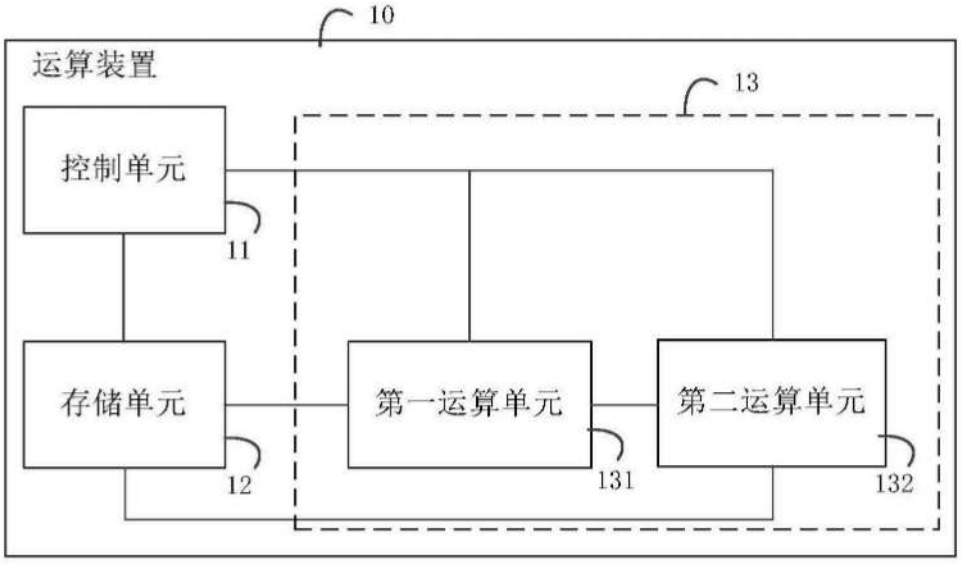

10
运算装置
13
控制单元
11
存储单元
12
第一运算单元
131
第二运算单元
132

图2

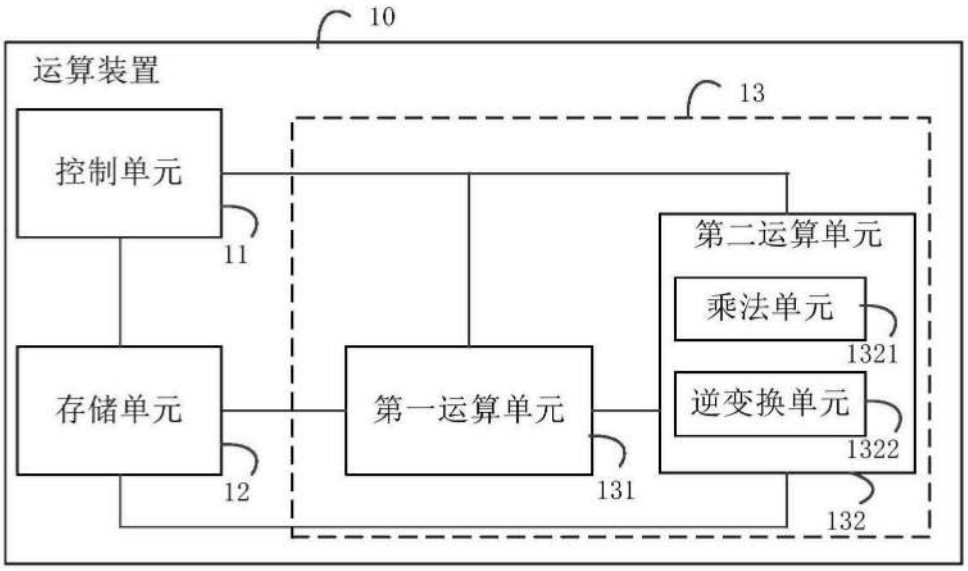

10
运算装置
13
控制单元
11
第二运算单元
乘法单元
1321
存储单元
第一运算单元
逆变换单元
1322
12
131
132

图3

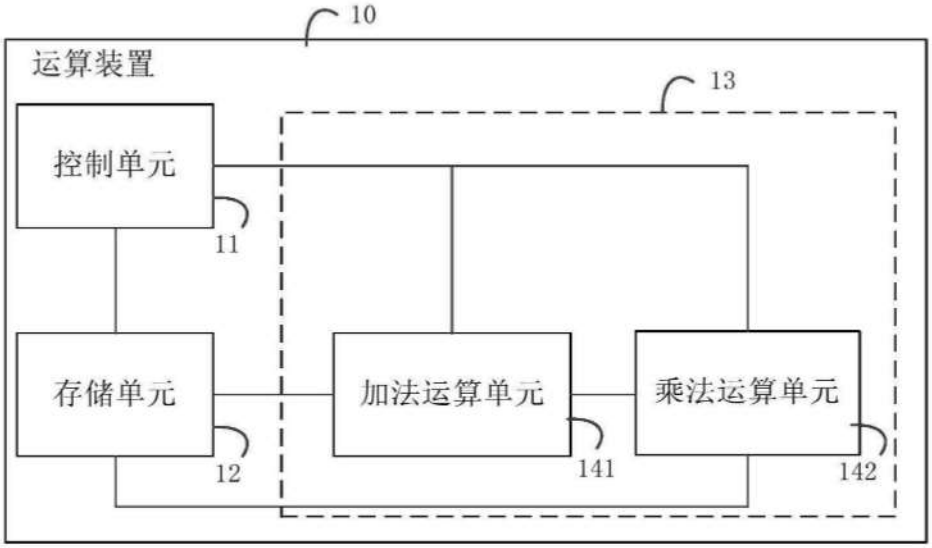

10
运算装置
13
控制单元
11
存储单元
加法运算单元
乘法运算单元
12
141
142

图4

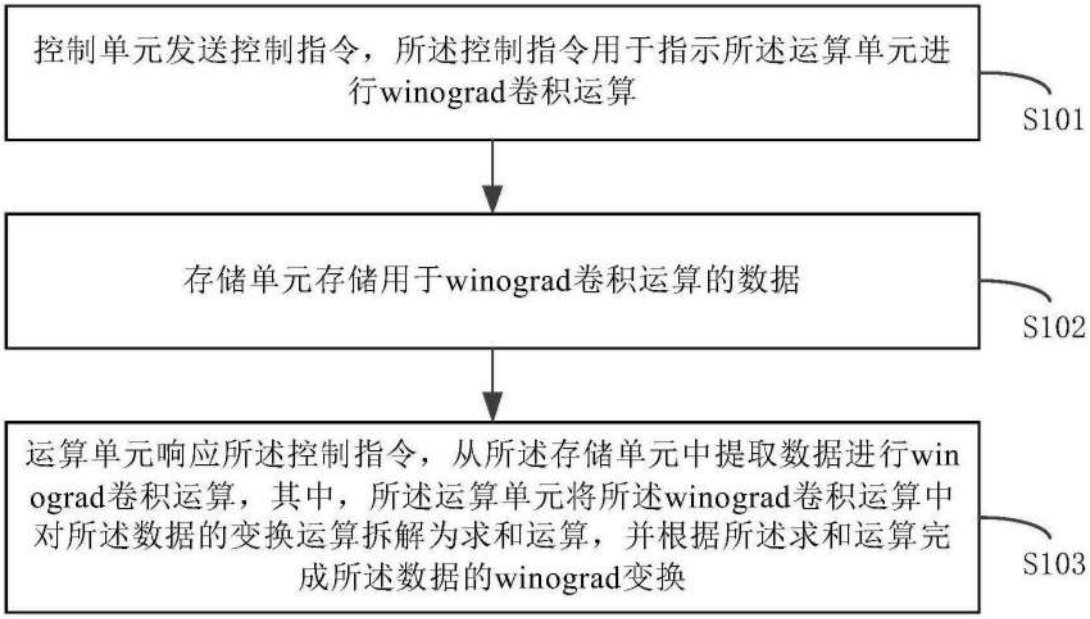

控制单元发送控制指令，所述控制指令用于指示所述运算单元进行winograd卷积运算
S101
存储单元存储用于winograd卷积运算的数据
S102
运算单元响应所述控制指令，从所述存储单元中提取数据进行winograd卷积运算，其中，所述运算单元将所述winograd卷积运算中对所述数据的变换运算拆解为求和运算，并根据所述求和运算完成所述数据的winograd变换
S103

图5

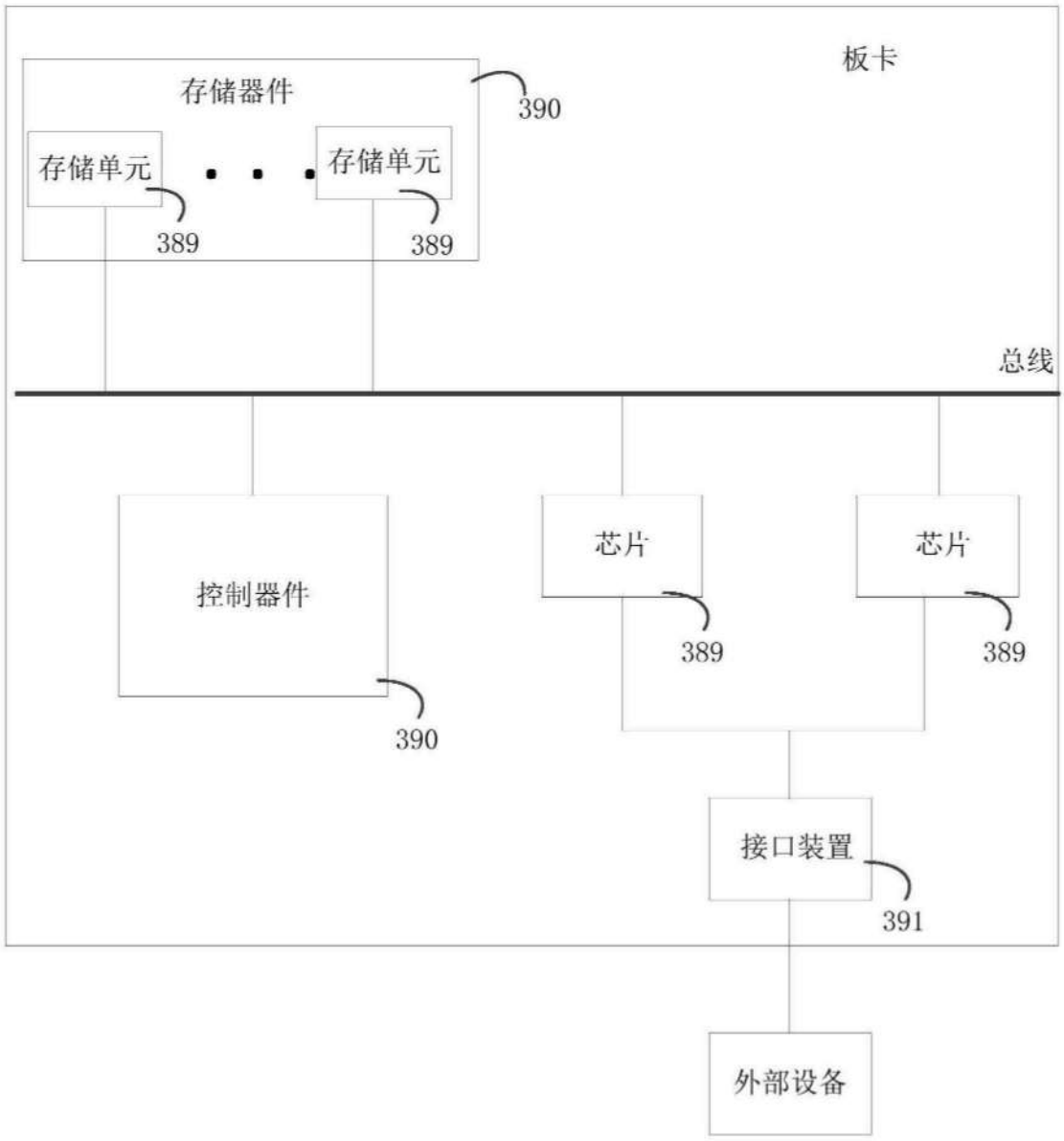

板卡
存储器件
390
存储单元
存储单元
389
389
总线
控制器件
390
芯片
389
芯片
389
接口装置
391
外部设备

图6