(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 112766472 A
(43)申请公布日 2021.05.07

(21)申请号 201911060473.X

(22)申请日 2019.11.01

(71)申请人 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 不公告发明人

(74)专利代理机构 北京林达刘知识产权代理事务所(普通合伙) 11277
代理人 刘新宇

(51)Int.Cl.
G06N 3/063(2006.01)
G06T 1/20(2006.01)
G06F 9/30(2006.01)

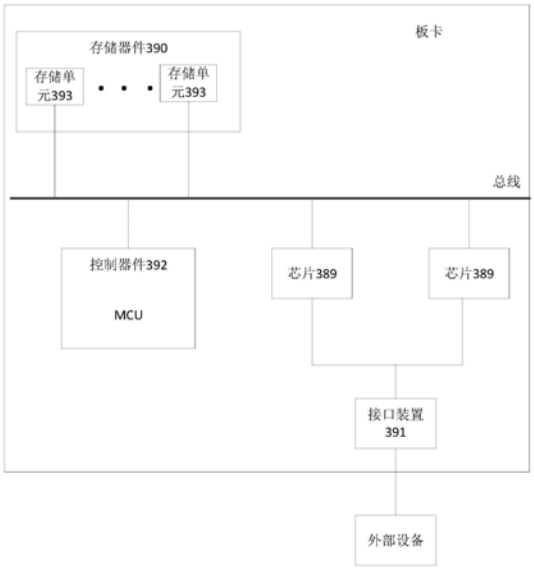
权利要求书3页 说明书29页 附图6页

(54)发明名称

数据处理方法、装置、计算机设备和存储介质

(57)摘要

本公开涉及一数据处理方法、装置、计算机设备和存储介质。其所公开的板卡包括:存储器件、接口装置和控制器件以及包括数据处理装置的人工智能芯片;其中,人工智能芯片与存储器件、控制器件以及接口装置分别连接;存储器件,用于存储数据;接口装置,用于实现人工智能芯片与外部设备之间的数据传输;控制器件,用于对人工智能芯片的状态进行监控。本公开实施例所提供的数据处理方法、装置、计算机设备和存储介质,可以提高量化的精度的同时,节约winograd卷积的运算时间,减少能耗。



CN 112766472 A

1.一种数据处理方法,其特征在于,所述方法包括:

根据确定出的量化参数对待量化数据进行量化处理,得到量化后的数据,其中,所述量化参数是根据对应的待量化数据的统计结果和数据位宽确定的;

根据所述量化后的数据继续执行winograd卷积处理,得到量化后的winograd卷积结果;

对所述量化后的winograd卷积结果执行反量化处理,得到winograd卷积结果。

2.根据权利要求1所述的方法,其特征在于,所述winograd卷积处理过程,包括:

将待运算数据的winograd正变换拆解为求和运算,并进行计算得到所述待运算数据中每个数据的winograd正变换结果;

执行所述待运算数据中对应数据的winograd正变换结果之间的对位乘操作,得到对位乘结果;

将对所述对位乘结果的winograd逆变换拆解为求和运算,得到所述winograd卷积结果,

其中,所述待量化数据为所述待运算数据、所述待运算数据的winograd正变换结果和所述对位乘结果中的任一种。

3.根据权利要求2所述的方法,其特征在于,根据确定出的量化参数对待量化数据进行量化处理,得到量化后的数据,包括以下任一操作:

在将待运算数据的winograd正变换拆解为求和运算之前,将所述待运算数据作为待量化数据进行量化处理;

在进行对位乘操作之前,将待运算数据中每个数据的winograd正变换结果作为待量化数据进行量化处理;

在进行winograd逆变换拆解为求和运算之前,将所述对位乘结果作为待量化数据进行量化处理。

4.根据权利要求2所述的方法,其特征在于,所述将待运算数据的winograd正变换拆解为求和运算,并进行计算得到所述待运算数据中每个数据的winograd正变换结果,包括:

将所述待运算数据中的每个数据分别拆解为多个第一子张量,对所述待运算数据中的每个数据的多个第一子张量进行winograd正变换并求和得到所述待运算数据中每个数据的winograd正变换结果,

其中,每个数据拆分出的多个第一子张量的个数与对应的数据中不为0元素的个数相同,每个第一子张量中有一个元素与对应数据中的对应位置的元素相同、其他元素均为0。

5.根据权利要求2所述的方法,其特征在于,所述将对所述对位乘结果的winograd逆变换拆解为求和运算,得到所述winograd卷积结果,包括:

将所述对位乘结果拆解为多个第二子张量,对所述多个第二子张量进行winograd逆变换并求和,得到所述待运算数据的winograd卷积结果;

其中,所述多个第二子张量的个数与所述对位乘结果中不为0元素的个数相同,所述多个第二子张量中的每个第二子张量中有一个元素与所述对位乘结果中的对应位置的元素相同、其他元素均为0。

6.根据权利要求1所述的方法,其特征在于,所述统计结果包括以下任一种:每种待量化数据中的绝对值最大值、每种待量化数据中的最大值和最小值的距离的二分之一,

所述量化参数包括点位置参数、缩放系数和偏移量中的一种或多种，

其中，所述绝对值最大值是每种待量化数据中的最大值或最小值的绝对值。

7.根据权利要求6所述的方法，其特征在于，所述缩放系数是根据所述点位置参数、所述统计结果、所述数据位宽确定的。

8.根据权利要求6所述的方法，其特征在于，所述偏移量是根据每种待量化数据的统计结果确定的。

9.根据权利要求6所述的方法，其特征在于，所述点位置参数是根据所述统计结果和所述数据位宽确定的。

10.根据权利要求6所述的方法，其特征在于，所述方法还包括：

根据所述数据位宽对应的量化误差，对所述数据位宽进行调整，以利用调整后的数据位宽确定量化参数，

其中，所述量化误差是根据对应层中量化后的数据与对应的量化前的数据确定的。

11.根据权利要求10所述的方法，其特征在于，所述量化前的数据是在目标迭代间隔内的权值更新迭代过程中涉及的待量化数据；

其中，所述目标迭代间隔包括至少一次权值更新迭代，且同一目标迭代间隔内量化过程中采用相同的所述数据位宽。

12.根据权利要求2所述的方法，其特征在于，所述待运算数据包括输入神经元、权值和梯度中的至少一种。

13.一种数据处理装置，其特征在于，所述装置包括：

数据量化模块，根据确定出的量化参数对待量化数据进行量化处理，得到量化后的数据，其中，所述量化参数是根据对应的待量化数据的统计结果和数据位宽确定的；

卷积处理模块，根据所述量化后的数据继续执行winograd卷积处理，得到量化后的winograd卷积结果；

反量化处理模块，对所述量化后的winograd卷积结果执行反量化处理，得到winograd卷积结果。

14.一种人工智能芯片，其特征在于，所述芯片包括如权利要求13所述的数据处理装置。

15.一种电子设备，其特征在于，所述电子设备包括如权利要求25所述的人工智能芯片。

16.一种板卡，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14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。

17.根据权利要求16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。

18.一种电子设备,其特征在于,包括:

处理器;

用于存储处理器可执行指令的存储器;

其中,所述处理器被配置为调用所述存储器存储的指令,以执行权利要求1至12中任意一项所述的方法。

19.一种计算机可读存储介质,其上存储有计算机程序指令,其特征在于,所述计算机程序指令被处理器执行时实现权利要求1至12中任意一项所述的方法。

数据处理方法、装置、计算机设备和存储介质

技术领域

[0001] 本公开涉及计算机技术领域，特别是涉及一种数据处理方法、装置、计算机设备和存储介质。

背景技术

[0002] 在人工智能技术领域，神经网络算法是最近非常流行的一种机器学习算法，在各种领域中都取得了非常好的效果，比如图像识别，语音识别，自然语言处理等。随着神经网络算法的发展，算法的复杂度也越来越高，为了提高识别度，模型的规模也在逐渐增大。用GPU和CPU处理起这些大规模的模型，要花费大量的计算时间，并且耗电量很大。

发明内容

[0003] 基于此，有必要针对上述技术问题，提供一种能够能够节约计算时间、减少能耗，提高计算精度的数据处理方法、装置、计算机设备和存储介质。

[0004] 根据本公开的一方面，提供了一种数据处理方法，所述方法包括：

[0005] 根据确定出的量化参数对待量化数据进行量化处理，得到量化后的数据，其中，所述量化参数是根据对应的待量化数据的统计结果和数据位宽确定的；

[0006] 根据所述量化后的数据继续执行winograd卷积处理，得到量化后的winograd卷积结果；

[0007] 对所述量化后的winograd卷积结果执行反量化处理，得到winograd卷积结果。

[0008] 根据本公开的另一方面，提供了一种数据处理装置，所述装置包括：

[0009] 数据量化模块，根据确定出的量化参数对待量化数据进行量化处理，得到量化后的数据，其中，所述量化参数是根据对应的待量化数据的统计结果和数据位宽确定的；

[0010] 卷积处理模块，根据所述量化后的数据继续执行winograd卷积处理，得到量化后的winograd卷积结果；

[0011] 反量化处理模块，对所述量化后的winograd卷积结果执行反量化处理，得到winograd卷积结果。

[0012] 根据本公开的另一方面，提供了一种人工智能芯片，所述芯片包括如前述任意一项所述的数据处理装置。

[0013] 根据本公开的另一方面，提供了一种电子设备，所述电子设备包括如前述的人工智能芯片。

[0014] 根据本公开的另一方面，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前述的人工智能芯片；

[0015] 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；

[0016] 所述存储器件，用于存储数据；

[0017] 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；

[0018] 所述控制器件，用于对所述人工智能芯片的状态进行监控。

[0019] 根据本公开的另一方面，提供了一种电子设备，包括：

[0020] 处理器；

[0021] 用于存储处理器可执行指令的存储器；

[0022] 其中，所述处理器被配置为调用所述存储器存储的指令，以执行前述中任意一项所述的方法。

[0023] 根据本公开的另一方面，提供了一种计算机可读存储介质，其上存储有计算机程序指令，其特征在于，所述计算机程序指令被处理器执行时实现前述中任意一项所述的方法。

[0024] 根据本公开的数据处理方法、装置、计算机设备和存储介质，根据确定出的量化参数对待量化数据进行量化处理，得到量化后的数据，其中，量化参数是根据对应的待量化数据的统计结果和数据位宽确定的；根据量化后的数据继续执行winograd卷积处理，得到量化后的winograd卷积结果；对量化后的winograd卷积结果执行反量化处理，得到winograd卷积结果。可以提高量化的精度的同时，节约winograd卷积的运算时间，减少能耗。

[0025] 根据下面参考附图对示例性实施例的详细说明，本公开的其它特征及方面将变得清楚。

附图说明

[0026] 包含在说明书中并且构成说明书的一部分的附图与说明书一起示出了本公开的示例性实施例、特征和方面，并且用于解释本公开的原理。

[0027] 图1示出根据本公开实施例的数据处理方法的流程图。

[0028] 图2示出根据本公开实施例的对称的定点数表示的示意图。

[0029] 图3示出根据本公开实施例的引入偏移量的定点数表示的示意图。

[0030] 图4a、图4b为训练过程中神经网络的权值数据变动幅度曲线图。

[0031] 图5示出根据本公开实施例的数据处理装置的框图。

[0032] 图6示出根据本公开实施例的板卡的结构框图。

[0033] 图7示出根据本公开实施例的一种电子设备800的框图。

[0034] 图8示出根据本公开实施例的一种电子设备1900的框图。

具体实施方式

[0035] 下面将结合本公开实施例中的附图，对本公开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公开一部分实施例，而不是全部的实施例。基于本公开中的实施例，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，都属于本公开保护的范围。

[0036] 应当理解，本公开的权利要求、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。本公开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。

[0037] 还应当理解，在此本公开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,而并不意在限定本公开。如在本公开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,在本公开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。

[0038] 如在本说明书和权利要求书中所使用的那样,术语“如果”可以依据上下文被解释为“当...时”或“一旦”或“响应于确定”或“响应于检测到”。类似地,短语“如果确定”或“如果检测到[所描述条件或事件]”可以依据上下文被解释为意指“一旦确定”或“响应于确定”或“一旦检测到[所描述条件或事件]”或“响应于检测到[所描述条件或事件]”。

[0039] 根据本公开实施例的数据处理方法可应用于处理器中,该处理器可以是通用处理器,例如CPU(Central Processing Unit,中央处理器),也可以是用于执行人工智能运算的人工智能处理器(IPU)。人工智能运算可包括机器学习运算,类脑运算等。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。该人工智能处理器可例如包括GPU(Graphics Processing Unit,图形处理单元)、NPU(Neural-Network Processing Unit,神经网络处理单元)、DSP(Digital Signal Process,数字信号处理单元)、现场可编程门阵列(Field－Programmable Gate Array,FPGA)芯片中的一种或组合。本公开对处理器的具体类型不作限制。

[0040] 在一种可能的实现方式中,本公开中所提及的处理器可包括多个处理单元,每个处理单元可以独立运行所分配到的各种任务,如:卷积运算任务、池化任务或全连接任务等。本公开对处理单元及处理单元所运行的任务不作限制。

[0041] 处理器包括多个处理单元以及存储单元,多个处理单元用于执行指令序列,存储单元用于存储数据,可包括随机存储器(RAM,Random Access Memory)和寄存器堆。处理器中的多个处理单元既可共用部分存储空间,例如共用部分RAM存储空间和寄存器堆,又可同时拥有各自的存储空间。

[0042] 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:神经元、权值进行一定规模切分后分别进行线性变换(winograd正变换),再将变换后的神经元和权值进行对位乘法,最后对对位乘法结果再次进行线性变换(winograd逆变换)得到与原卷积操作等价的卷积结果。

[0043] winograd变换的表达式如下所示:

[0044] 对于一维的神经元和权值:$S=A^T((Gg)\odot(B^Td))$

[0045] 对于二维的神经元和权值:$S=A^T((GgG^T)\odot(B^TdB))A$

[0046] 其中,g表示权值,G表示权值对应的左乘正变换矩阵,G^T表示权值对应的右乘正变换矩阵,d表示输入神经元,B表示输入神经元对应的右乘正变换矩阵,B^T表示输入神经元对应的左乘正变换矩阵,⊙表示对位乘运算,A表示右乘逆变换矩阵,A^T表示左乘逆变换矩阵。对于不同维度的输入神经元,都有与其相对应的B和B^T;同样的,对于不同维度的权值,都有与其相对应的G和G^T。

[0047] 通过winograd卷积替代原始卷积操作能够带来硬件能效比和运算时间上的较大收益,同时也可以在不增加、或者增加较少的硬件开销的情况下实现更高的神经网络性能。但是,winograd卷积的弊端仍然较为明显,大量的乘法运算在计算过程中仍然消耗较长的运算时间。

[0048] 为了解决上述技术问题，本公开提供了一种数据处理方法、装置、计算机设备和存储介质，可以将winograd卷积过程中的乘法运算拆解为加法运算，从而节约计算时间、减少能耗，并且对winograd卷积过程中的数据进行量化处理，进一步的提高计算性能。

[0049] 图1示出根据本公开实施例的数据处理方法的流程图。如图1所示，该方法应用于处理器，该方法包括步骤S11至步骤S13。

[0050] 在步骤S11中，根据确定出的量化参数对待量化数据进行量化处理，得到量化后的数据，其中，所述量化参数是根据对应的待量化数据的统计结果和数据位宽确定的。

[0051] 在一种可能的实现方式中，该方法还可以包括：对待量化数据进行统计，确定每种待量化数据的统计结果；利用每种待量化数据的统计结果以及数据位宽确定对应量化数据。

[0052] 在一种可能的实现方式中，所述统计结果可以包括以下任一种：每种待量化数据中的绝对值最大值、每种待量化数据中的最大值和最小值的距离的二分之一，

[0053] 所述量化参数可以包括点位置参数、缩放系数和偏移量中的一种或多种，

[0054] 其中，所述绝对值最大值是每种待量化数据中的最大值或最小值的绝对值。

[0055] 在一种可能的实现方式中，所述缩放系数是根据所述点位置参数、所述统计结果、所述数据位宽确定的。

[0056] 在一种可能的实现方式中，所述偏移量是根据每种待量化数据的统计结果确定的。

[0057] 在一种可能的实现方式中，所述点位置参数是根据所述统计结果和所述数据位宽确定的。

[0058] 在一种可能的实现方式中，量化参数可以分以下六种情况。第一种情况：量化参数是点位置参数s。这种情况下，可以利用如下的公式(1)对待量化数据进行量化，得到量化数据I_x：

[0059] $$I_x = round\left(\frac{F_x}{2^s}\right) \quad (1)$$

[0060] 其中，s为点位置参数，I_x为数据x量化后的n位二进制表示值，F_x为数据x量化前的浮点值，round为进行四舍五入的取整运算。需要说明的是，此处不仅仅局限于round这一种取整运算，也可以采用其他的取整运算方法，例如：采用向上取整、向下取整、向零取整等取整运算，替换公式(1)中的round取整运算。此时，用n位定点数可以表示浮点数的最大值A为$2^s(2^{n-1}-1)$，那么n位定点数可以表示待量化数据的数域中最大值为$2^s(2^{n-1}-1)$，n位定点数可以表示待量化数据的数域中最小值为$-2^s(2^{n-1}-1)$。由式(1)可知，采用第一种情况对应的量化参数对待量化数据进行量化时，量化间隔为2^s，量化间隔记为C。

[0061] 设Z为待量化数据的数域中所有浮点数的绝对值最大值，则A需要包含Z，且Z要大于$\frac{A}{2}$，因此有如下公式(2)约束：

[0062] $$2^s(2^{n-1}-1) \geq Z > 2^{s-1}(2^{n-1}-1) \quad (2)$$

[0063] 因此，$\log_2\left(\frac{Z}{2^{n-1}-1}\right)-1 > s \geq \log_2\left(\frac{Z}{2^{n-1}-1}\right)$，得到$s = ceil\left(\log_2\left(\frac{Z}{2^{n-1}-1}\right)\right)$，

$A=2^{ceil\left(\log_2\frac{Z}{2^{n-1}-1}\right)}\left(2^{n-1}-1\right)$。

[0064] 根据式(3)对数据x量化后的n位二进制表示值I_x进行反量化，获得反量化数据$\hat{F}_x$；其中，所述反量化数据$\hat{F}_x$的数据格式与对应的量化前的数据F_x的数据格式相同，均为浮点值。

[0065] $$\hat{F}_x=round(\frac{F_x}{2^s})\times 2^s \qquad (3)$$

[0066] 第二种情况：量化参数是第一缩放系数f_1。这种情况下，可以利用如下的公式(4)对待量化数据进行量化，得到量化数据I_x：

[0067] $$I_x=round\left(\frac{F_x}{f_1}\right) \qquad (4)$$

[0068] 其中，f_1为第一缩放系数，I_x为数据x量化后的n位二进制表示值，F_x为数据x量化前的浮点值，round为进行四舍五入的取整运算。需要说明的是，此处不仅仅局限于round这一种取整运算，也可以采用其他的取整运算方法，例如：采用向上取整、向下取整、向零取整等取整运算，替换公式(4)中的round取整运算。由式(4)可知，采用第二种情况对应的量化参数对待量化数据进行量化时，量化间隔为f_1，量化间隔记为C。

[0069] 对于第一缩放系数f_1来说，有一种情况，即：点位置参数s为固定已知值，不再发生变化，设$2^s=T$，T为固定值，那么，用n位定点数可以表示浮点数的最大值A为$(2^{n-1}-1)\times T$。这种情况下，最大值A取决于数据位宽n。设Z为待量化数据的数域中所有数的绝对值最大值，则$f_1=\frac{Z}{2^{n-1}-1}$，此时$Z=(2^{n-1}-1)\times f_1$。n位定点数可以表示待量化数据的数域中最大值为$(2^{n-1}-1)\times f_1$，n位定点数可以表示待量化数据的数域中最小值为$-(2^{n-1}-1)\times f_1$。还有一种情况，在工程应用中，$2^s\times f_2$作为一个整体当做第一缩放系数f_1。此时，就可以当做不存在独立的点位置参数s。其中，f_2为第二缩放系数。设Z为待量化数据的数域中所有数的绝对值最大值，则$f_1=\frac{Z}{2^{n-1}-1}$，此时$Z=(2^{n-1}-1)\times f_1$。n位定点数可以表示待量化数据的数域中最大值为$(2^{n-1}-1)\times f_1$，n位定点数可以表示待量化数据的数域中最小值为$-(2^{n-1}-1)\times f_1$。

[0070] 根据式(5)对数据x量化后的n位二进制表示值I_x进行反量化，获得反量化数据$\hat{F}_x$；其中，所述反量化数据$\hat{F}_x$的数据格式与对应的量化前的数据F_x的数据格式相同，均为浮点值。

[0071] $$\hat{F}_x=round(\frac{F_x}{f_1})\times f_1 \qquad (5)$$

[0072] 第三种情况：量化参数是点位置参数s和第二缩放系数f_2。这种情况下，可以利用如下的公式(6)对待量化数据进行量化，得到量化数据I_x：

[0073] $$I_x=round\left(\frac{F_x}{2^s\times f_2}\right) \qquad (6)$$

[0074] 其中，s为点位置参数，f_2为第二缩放系数，$f_2=\frac{Z}{2^s\left(2^{n-1}-1\right)}$；$I_x$为数据x量化后的n位二进制表示值，$F_x$为数据x量化前的浮点值，round为进行四舍五入的取整运算。需要说明的是，此处不仅仅局限于round这一种取整运算，也可以采用其他的取整运算方法，例如：采用向上取整、向下取整、向零取整等取整运算，替换公式(6)中的round取整运算。用n位定点数可以表示的待量化数据的数域中的最大值A为$2^s(2^{n-1}-1)$。由式(6)可知，采用第三种情况对应的量化参数对待量化数据进行量化时，量化间隔为$2^s\times f_2$，量化间隔记为C。

[0075] 设Z为待量化数据的数域中所有数的绝对值最大值，此时，根据公式(2)可得：

[0076] $1\geq\frac{Z}{2^s\left(2^{n-1}-1\right)}>\frac{1}{2}$，即$1\geq\frac{Z}{A}>\frac{1}{2}$，$1\geq f_2>\frac{1}{2}$

[0077] $f_2=\frac{Z}{2^s\left(2^{n-1}-1\right)}=\frac{Z}{A}$时，根据公式(2)，Z可以无损精确表示。当$f_2=1$时，公式(6)与公式(1)，$s=ceil\left(\log_2\left(\frac{Z}{2^{n-1}-1}\right)\right)$。n位定点数可以表示待量化数据的数域中最大值为$(2^{n-1}-1)\times2^s\times f_2$，n位定点数可以表示待量化数据的数域中最小值为$-(2^{n-1}-1)\times2^s\times f_2$。

[0078] 根据式(7)对数据x量化后的n位二进制表示值I_x进行反量化，获得反量化数据$\hat{F}_x$；其中，所述反量化数据$\hat{F}_x$的数据格式与对应的量化前的数据F_x的数据格式相同，均为浮点值。

[0079] $$\hat{F}_x=round\left(\frac{F_x}{2^s\times f_2}\right)\times2^s\times f_2 \qquad (7)$$

[0080] 在本实施例中，图2示出根据本公开实施例的对称的定点数表示的示意图。如图2所示的待量化数据的数域是以“0”为对称中心分布。Z为待量化数据的数域中所有浮点数的绝对值最大值，在图2中，A为n位定点数可以表示的浮点数的最大值，浮点数A转换为定点数是$2^{n-1}-1$。为了避免溢出，A需要包含Z。在实际运算中，神经网络运算过程中的浮点数据趋向于某个确定区间的正态分布，但是并不一定满足以“0”为对称中心的分布，这时用定点数表示时，容易出现溢出情况。为了改善这一情况，量化参数中引入偏移量。图3示出根据本公开实施例的引入偏移量的定点数表示的示意图。如图3所示。待量化数据的数域不是以“0”为对称中心分布，Z_{min}是待量化数据的数域中所有浮点数的最小值，Z_{max}是待量化数据的数域中所有浮点数的最大值。P为Z_{min}～Z_{max}之间的中心点，将待量化数据的数域整体偏移，使得平移后的待量化数据的数域以“0”为对称中心分布，平移后的待量化数据的数域中的绝对值最大值为Z。由图3可知，偏移量为“0”点到“P”点之间的水平距离，该距离称为偏移量O。其中，$O=\frac{Z_{min}+Z_{max}}{2}$，$Z=\frac{Z_{max}-Z_{min}}{2}$。

[0081] 基于上述关于偏移量O的描述，出现第四种量化参数的情况。第四种情况：量化参数包括点位置参数和偏移量。这种情况下，可以利用如下的公式(8)对待量化数据进行量化，得到量化数据I_x：

[0082] $$I_x=round\left(\frac{F_x-O}{2^s}\right) \quad (8)$$

[0083] 其中，s为点位置参数，0为偏移量，$O=\frac{Z_{min}+Z_{max}}{2}$，$I_x$为数据x量化后的n位二进制表示值，$F_x$为数据x量化前的浮点值，round为进行四舍五入的取整运算。需要说明的是，此处不仅仅局限于round这一种取整运算，也可以采用其他的取整运算方法，例如：采用向上取整、向下取整、向零取整等取整运算，替换公式(8)中的round取整运算。此时，用n位定点数可以表示浮点数的最大值A为$2^s(2^{n-1}-1)$，那么n位定点数可以表示待量化数据的数域中最大值为$2^s(2^{n-1}-1)+0$，n位定点数可以表示待量化数据的数域中最小值为$-2^s(2^{n-1}-1)+0$。由式(8)可知，采用第四种情况对应的量化参数对待量化数据进行量化时，量化间隔为2^s，量化间隔记为C。

[0084] 设Z为待量化数据的数域中所有浮点数的绝对值最大值，$Z=\frac{Z_{max}-Z_{min}}{2}$，则A需要包含Z，且Z要大于$\frac{A}{2}$，根据公式(2)获得$\log_2\left(\frac{Z}{2^{n-1}-1}\right)-1>s\geq\log_2\left(\frac{Z}{2^{n-1}-1}\right)$，进而得到$s=ceil\left(\log_2\left(\frac{Z}{2^{n-1}-1}\right)\right)$，$A=2^{ceil\left(\log_2\frac{Z}{2^{n-1}-1}\right)}\left(2^{n-1}-1\right)$。

[0085] 根据式(9)对数据x量化后的n位二进制表示值I_x进行反量化，获得反量化数据$\hat{F}_x$；其中，所述反量化数据$\hat{F}_x$的数据格式与对应的量化前的数据F_x的数据格式相同，均为浮点值。

[0086] $$\hat{F}_x=round(\frac{F_x-O}{2^s})\times 2^s+O \quad (9)$$

[0087] 基于上述关于偏移量0的描述，出现第五种量化参数的情况。第五种情况：量化参数包括第一缩放系数f_1和偏移量0。这种情况下，可以利用如下的公式(10)对待量化数据进行量化，得到量化数据I_x：

[0088] $$I_x=round\left(\frac{F_x-O}{f_1}\right) \quad (10)$$

[0089] 其中，f_1为第一缩放系数，0为偏移量，I_x为数据x量化后的n位二进制表示值，F_x为数据x量化前的浮点值，round为进行四舍五入的取整运算。需要说明的是，此处不仅仅局限于round这一种取整运算，也可以采用其他的取整运算方法，例如：采用向上取整、向下取整、向零取整等取整运算，替换公式(10)中的round取整运算。此时，有一种情况，即：点位置参数s为固定已知值，不再发生变化，设$2^s=T$，T为固定值。那么，用n位定点数可以表示浮点数的最大值A为$(2^{n-1}-1)\times T$。这种情况下，最大值A取决于数据位宽n。设Z为待量化数据的数域中所有数的绝对值最大值，则$f_1=\frac{Z}{2^{n-1}-1}$，此时$Z=(2^{n-1}-1)\times f_1$。n位定点数可以表示待量化数据的数域中最大值为$(2^{n-1}-1)\times f_1$，n位定点数可以表示待量化数据的数域中最小值为$-(2^{n-1}-1)\times f_1$。还有一种情况，在工程应用中，$2^s\times f_2$作为一个整体当做第一缩放系数f_1。

此时，就可以当做不存在独立的点位置参数s。其中，f_2为第二缩放系数。设Z为待量化数据的数域中所有数的绝对值最大值，则$f_1=\frac{Z}{2^{n-1}-1}$，此时Z＝$(2^{n-1}-1)\times f_1$。n位定点数可以表示待量化数据的数域中最大值为$(2^{n-1}-1)\times f_1+0$，n位定点数可以表示待量化数据的数域中最小值为$-(2^{n-1}-1)\times f_1+0$。

[0090] 由式(10)可知，采用第五种情况对应的量化参数对待量化数据进行量化时，量化间隔为f_1，量化间隔记为C。

[0091] 根据式(11)对数据x量化后的n位二进制表示值I_x进行反量化，获得反量化数据$\hat{F}_x$；其中，所述反量化数据$\hat{F}_x$的数据格式与对应的量化前的数据F_x的数据格式相同，均为浮点值。

[0092] $$\hat{F}_x=round(\frac{F_x-O}{f_1})\times f_1+O \qquad (11)$$

[0093] 基于上述关于偏移量0的描述，出现第六种量化参数的情况。第六种情况：量化参数包括点位置参数、第二缩放系数f_2和偏移量0。这种情况下，可以利用如下的公式(12)对待量化数据进行量化，得到量化数据I_x：

[0094] $$I_x=round\left(\frac{F_x-O}{2^s\times f_2}\right) \qquad (12)$$

[0095] 其中，s为点位置参数，偏移量0，f_2为第二缩放系数，$f_2=\frac{Z}{2^s\left(2^{n-1}-1\right)}$；$Z=\frac{Z_{\max}-Z_{\min}}{2}$；$I_x$为数据x量化后的n位二进制表示值，$F_x$为数据x量化前的浮点值，round为进行四舍五入的取整运算。需要说明的是，此处不仅仅局限于round这一种取整运算，也可以采用其他的取整运算方法，例如：采用向上取整、向下取整、向零取整等取整运算，替换公式(12)中的round取整运算。用n位定点数可以表示的待量化数据的数域中的最大值A为$2^s(2^{n-1}-1)$。由式(12)可知，采用第六种情况对应的量化参数对待量化数据进行量化时，量化间隔为$2^s\times f_2$，量化间隔记为C。

[0096] 设Z为待量化数据的数域中所有数的绝对值最大值，此时，根据公式(2)可得：

[0097] $1\geq\frac{Z}{2^s\left(2^{n-1}-1\right)}>\frac{1}{2}$，即$1\geq\frac{Z}{A}>\frac{1}{2}$，$1\geq f_2>\frac{1}{2}$

[0098] $f_2=\frac{Z}{2^s\left(2^{n-1}-1\right)}=\frac{Z}{A}$时，根据公式(2)，Z可以无损精确表示。当$f_2=1$时，$s=ceil\left(\log_2\left(\frac{Z_{\max}-Z_{\min}}{2(2^{n-1}-1)}\right)\right)$。n位定点数可以表示待量化数据的数域中最大值为$(2^{n-1}-1)\times 2^s\times f_2+0$，n位定点数可以表示待量化数据的数域中最小值为$-(2^{n-1}-1)\times 2^s\times f_2+0$。

[0099] 根据式(13)对数据x量化后的n位二进制表示值I_x进行反量化，获得反量化数据$\hat{F}_x$；其中，所述反量化数据$\hat{F}_x$的数据格式与对应的量化前的数据F_x的数据格式相同，均为浮点值。

[0100] $$\hat{F}_x = round\left(\frac{F_x}{2^s \times f_2}\right) \times 2^s \times f_2 + O \qquad (13)$$

[0101] 在本实施例中，由公式(1)～公式(13)可知，点位置参数和缩放系数均与数据位宽有关。不同的数据位宽，导致点位置参数和缩放系数不同，从而影响量化精度。量化就是将以往用32bit或者64bit表达的高精度数转换成占用较少内存空间的定点数的过程，高精度数转换为定点数的过程就会在精度上引起一定的损失。在训练或微调过程中，在一定的迭代(iterations)的次数范围内，使用相同的数据位宽量化对神经网络运算的总体精度影响不大。超过一定的迭代次数，再使用同一数据位宽量化就无法满足训练或微调对精度的要求。这就需要随着训练或微调的过程对数据位宽n进行调整。简单地，可以人为将数据位宽n设置为预设值。在不同的迭代次数范围内，调用提前设置的对应的数据位宽n。

[0102] 在一种可能的实现方式中，所述方法还可以包括：

[0103] 根据所述数据位宽对应的量化误差，对所述数据位宽进行调整，以利用调整后的数据位宽确定量化参数，

[0104] 其中，所述量化误差是根据对应层中量化后的数据与对应的量化前的数据确定的。

[0105] 在一种可能的实现方式中，根据数据位宽对应的量化误差，对数据位宽进行调整，可以包括：对量化误差与阈值进行比较，根据比较结果调整数据位宽。其中，阈值可以包括第一阈值和第二阈值中的至少一个。第一阈值大于第二阈值。

[0106] 在一种可能的实现方式中，对量化误差与阈值进行比较，根据比较结果调整数据位，可以包括以下任一项：

[0107] 在量化误差大于或等于第一阈值时，增加数据位宽；

[0108] 在量化误差小于或等于第二阈值时，减少数据位宽；

[0109] 在量化误差处于第一阈值和第二阈值之间时，数据位宽保持不变。

[0110] 在该实现方式中，第一阈值和第二阈值可以为经验值，也可以为可变的超参数。常规的超参数的优化方法均适于第一阈值和第二阈值，这里不再赘述超参数的优化方案。

[0111] 需要强调的是，可以将数据位宽按照固定的位数步长进行调整，也可以根据量化误差与误差阈值之间的差值的不同，按照可变的调整步长调整数据位宽，最终根据神经网络运算过程的实际需要，将数据位宽调整的更长或更短。比如：当前卷积层的数据位宽n为16，根据量化误差将数据位宽n调整为12。也就是说，在实际应用中，数据位宽n取值为12而不必取值为16即可满足神经网络运算过程中对精度的需求，这样在精度允许范围内可以大大提到定点运算速度，从而提升了人工智能处理器芯片的资源利用率。

[0112] 在一种可能的实现方式中，该方法还可以包括：对量化后的数据进行反量化，获得反量化数据，其中，反量化数据的数据格式与对应的量化前的数据的数据格式相同；根据量化后的数据以及对应的反量化数据确定量化误差。

[0113] 在一种可能的实现方式中，量化前的数据可以是待量化数据。

[0114] 在一种可能的实现方式中，处理器可以根据待量化数据及其对应的反量化数据计算获得量化误差。设待量化数据为$Z=[z_1, z_2 \cdots, z_m]$，该待量化数据对应的反量化数据为$Z^{(n)}=[z_1^{(n)}, z_2^{(n)} \cdots, z_m^{(n)}]$。处理器可以根据该待量化数据Z及其对应的翻反量化数据$Z^{(n)}$确定误差项，并根据该误差项确定量化误差。

[0115] 在一种可能的实现方式中，处理器可以分别计算待量化数据Z与对应的反量化数据$Z^{(n)}$的差值，获得m个差值，并将该m个差值的和作为误差项。之后，处理器可以根据该误差项确定量化误差。具体的量化误差可以按照如下公式确定：

[0116] $$diff_{bit} = \log_2\left(\frac{\sum_i \left|Z_i^{(n)} - Z_i\right|}{\sum_i |Z_i|} + 1\right) \quad (14)$$

[0117] 其中，i为待量化数据集合中第i个待量化数据的下标。i为大于或等于1、且小于或等于m的整数。

[0118] 应当理解的是，上述量化误差的确定方式仅是本公开的一个示例，本领域技术人员可以根据实际需要对量化误差的确定方式进行设置，本公开对此不作限制。

[0119] 对于数据位宽来说，图4a、图4b为训练过程中神经网络的权值数据变动幅度曲线图。在图4a和图4b中，横坐标表示是迭代数，纵坐标表示是权值取对数后的最大值。图4a所示的权值数据变动幅度曲线展示神经网络的任一卷积层同一周期(epoch)内在不同迭代对应的权值数据变动情况。在图4b中，conv0层对应权值数据变动幅度曲线A，conv1层对应权值数据变动幅度曲线B，conv2层对应权值数据变动幅度曲线C，conv3层对应权值数据变动幅度曲线D，conv4层对应权值数据变动幅度曲线e。由图4a和图4b可知，同一个周期(epoch)内，在训练初期，每次迭代权值变化幅度比较大。在训练中后期，每次迭代权值的变化幅度不会太大。此种情况下，在训练中后期，因为每次迭代前后权值数据变化幅度不大，使得每代的对应层的权值数据之间在一定的迭代间隔内具有相似性，在神经网络训练过程中每层涉及的数据量化时可以采用上一迭代时对应层量化时使用的数据位宽。但是，在训练初期，由于每次迭代前后权值数据的变化幅度比较大，为了满足量化所需的浮点运算的精度，在训练初期的每一次迭代，利用上一代对应层量化时采用的数据位宽对当前代的对应层的权值数据进行量化，或者基于当前层预设的数据位宽n对当前层的权值数据进行量化，获得量化后的定点数。根据量化后的权值数据和对应的量化前的权值数据，确定量化误差，根据量化误差与阈值的比较结果，对上一代对应层量化时采用的数据位宽或者当前层预设的数据位宽进行调整，将调整后的数据位宽应用于当前代的对应层的权值数据的量化。进一步地，在训练或微调过程中，神经网络的每层之间的权值数据相互独立，不具备相似性。因权值数据不具备相似性使得每层之间的神经元数据也相互独立，不具备相似性。因此，在神经网络训练或微调过程中，神经网络的每一迭代内的每层的数据位宽应用于对应层。上述以权值数据为例，在神经网络训练或微调过程中，神经元数据和梯度数据分别对应的数据位宽亦如此，此处不再赘述。

[0120] 在一种可能的实现方式中，所述量化前的数据是在目标迭代间隔内的权值更新迭代过程中涉及的待量化数据。其中，所述目标迭代间隔包括至少一次权值更新迭代，且同一目标迭代间隔内量化过程中采用相同的所述数据位宽。

[0121] 在一种可能的实现方式中，目标迭代间隔是根据在预判时间点权值更新迭代过程中涉及的待量化数据的点位置参数的变化趋势值确定的。或者目标迭代间隔是根据在预判时间点权值更新迭代过程中涉及的待量化数据的点位置参数的变化趋势值和数据位宽的变化趋势值确定的。其中，预判时间点是用于判断是否需要对数据位宽进行调整的时间点，预判时间点对应权值更新迭代完成时的时间点。

[0122] 在一种可能的实现方式中，目标迭代间隔的确定步骤可以包括：

[0123] 在预判时间点，确定权值迭代过程中待量化数据对应点位置参数的变化趋势值；

[0124] 根据点位置参数的变化趋势值确定对应目标迭代间隔。

[0125] 在该实现方式中，按照式(15)，点位置参数的变化趋势值根据当前预判时间点对应的权值迭代过程中的点位置参数的滑动平均值、上一预判时间点对应的权值迭代过程中的点位置参数的滑动平均值确定，或者根据当前预判时间点对应的权值迭代过程中的点位置参数、上一预判时间点对应的权值迭代过程中的点位置参数的滑动平均值确定。公式(15)的表达式为：

[0126] $diff_{update1}=|M^{(t)}-M^{(t-1)}|=\alpha|s^{(t)}-M^{(t-1)}|$ (15)

[0127] 式15中，M为点位置参数s随着训练迭代增加的滑动平均值。其中，$M^{(t)}$为第t个预判时间点对应的点位置参数s随着训练迭代增加的滑动平均值，根据公式(16)获得$M^{(t)}$。$s^{(t)}$为第t个预判时间点对应的点位置参数s。$M^{(t-1)}$为第t-1个预判时间点对应的点位置参数s的滑动平均值，α为超参数。$diff_{update1}$衡量点位置参数s变化趋势，由于点位置参数s的变化也变相体现在当前待量化数据中数据最大值Z_{max}的变化情况。$diff_{update1}$越大，说明数值范围变化剧烈，需要间隔更短的更新频率，即目标迭代间隔更小。

[0128] $M^{(t)}\leftarrow\alpha\times s^{(t-1)}+(1-\alpha)\times M^{(t-1)}$ (16)

[0129] 在该实现方式中，根据式(17)确定目标迭代间隔。对于目标迭代间隔来说，同一目标迭代间隔内量化过程中采用相同的数据位宽，不同目标迭代间隔内量化过程中采用的数据位宽可以相同，也可以不同。

[0130] $$I=\frac{\beta}{diff_{update1}}-\gamma \quad (17)$$

[0131] 式(17)中，I为目标迭代间隔。β、γ为超参数。$diff_{update1}$为点位置参数的变化趋势值。

[0132] 在该实现方式中，预判时间点包括第一预判时间点，根据目标迭代间隔确定第一预判时间点。具体地，在训练或微调过程中的第t个预判时间点，利用上一代对应层量化时采用的数据位宽对当前代的对应层的权值数据进行量化，获得量化后的定点数，根据量化前的权值数据和对应的量化前的权值数据，确定量化误差。将量化误差分别与第一阈值和第二阈值进行比较，利用比较结果确定是否对上一代对应层量化时采用的数据位宽进行调整。假如：第t个第一预判时间点对应第100代，第99代使用的数据位宽为n_1。在第100代，根据数据位宽n_1确认量化误差，将量化误差与第一阈值、第二阈值进行比较，获得比较结果。如果根据比较结果确认数据位宽n_1无需改变，利用式(17)确认目标迭代间隔为8代，当第100代作为当前目标迭代间隔内的起始迭代，那么第100代～第107代作为当前目标迭代间隔，当第100代作为上一目标迭代间隔的最末迭代，那么第101代～第108代作为当前目标迭代间隔。在当前目标迭代间隔内量化时每代仍然延用上一个目标迭代间隔所使用的数据位宽n_1。这种情况，不同的目标迭代间隔之间量化时所使用的数据位宽可以相同。如果以第100代～第107代作为当前的目标迭代间隔，那么下一个目标迭代间隔内的第108代作为第t+1个第一预判时间点，如果第101代～第108代作为当前的目标迭代间隔，那么当前的目标迭代间隔内的第108代作为第t+1个第一预判时间点。在第t+1个第一预判时间点，根据数据位宽n_1确认量化误差，将量化误差与第一阈值、第二阈值进行比较，获得比较结果。根据比

较结果确定数据位宽n_1需要更改为n_2，并利用式(17)确认目标迭代间隔为55代。那么第108代～第163代或者第109代～第163代作为目标迭代间隔，在该目标迭代间隔内量化时每代使用数据位宽n_2。这种情况下，不同的目标迭代间隔之间量化时所使用的数据位宽可以不同。

[0133] 在该实现方式中，不管第一预判时间点是目标迭代间隔内的起始迭代还是最末迭代，均适于式(15)来获得点位置参数的变化趋势值。如果当前时刻的第一预判时间点为当前目标迭代间隔的起始迭代，那么在式(15)中，$M^{(t)}$为当前目标迭代间隔的起始迭代对应时间点所对应的点位置参数s随着训练迭代增加的滑动平均值，$s^{(t)}$为当前目标迭代间隔的起始迭代对应时间点所对应的点位置参数s，$M^{(t-1)}$为上一目标迭代间隔的起始迭代对应时间点所对应的点位置参数s随着训练迭代增加的滑动平均值。如果当前时刻的第一预判时间点为当前目标迭代间隔的最末迭代，那么在式(15)中，$M^{(t)}$为当前目标迭代间隔的最末迭代对应时间点所对应的点位置参数s随着训练迭代增加的滑动平均值，$s^{(t)}$为当前目标迭代间隔的最末迭代对应时间点所对应的点位置参数s，$M^{(t-1)}$为上一目标迭代间隔的最末迭代对应时间点所对应的点位置参数s随着训练迭代增加的滑动平均值。

[0134] 在该实现方式中，在包括第一预判时间点的基础上，预判时间点还可以包括第二预判时间点。第二预判时间点是根据数据变动幅度曲线确定的。基于大数据在神经网络训练过程中数据变动幅度情况，获得如图4a所示的数据变动幅度曲线。

[0135] 以权值数据为例，由图4a所示的数据变动幅度曲线可知，从训练开始到第T代的迭代间隔周期内，每次权值更新时，数据变动幅度非常大。在当前预判时间点，量化时，当前代先利用上一代的数据位宽n_1进行量化，获得的量化结果与对应的量化前的数据确定对应的量化误差，量化误差分别与第一阈值、第二阈值进行比较，根据比较结果对数据位宽n_1进行调整，获得数据位宽n_2。利用数据位宽n_2对当前代涉及的待量化权值数据进行量化。然后根据式(17)确定目标迭代间隔，从而确定第一预判时间点，在第一预判时间点再判断是否调整数据位宽以及如何调整，并根据公式(17)确定下一目标迭代间隔来获得下一个第一预判时间点。由于训练开始到第T代的迭代间隔周期内，每一次迭代前后权值数据变化幅度非常大，使得每代的对应层的权值数据之间不具有相似性，为了满足精度问题，量化时当前代的每层的数据不能延用上一代的对应层的对应量化参数，在前T代可以代代调整数据位宽，此时，量化时前T代中每代使用的数据位宽均不同，目标迭代间隔为1代。为了人工智能处理器芯片的资源达到最优化利用，前T代的目标迭代间隔可以根据图4a所示的数据变动幅度曲线图所揭示的规律提前预设好，即：根据数据变动幅度曲线前T代的目标迭代间隔直接预设，无需经过公式(17)确认前T代的每代对应的权值更新迭代完成时的时间点作为第二预判时间点。从而使得人工智能处理器芯片的资源更为合理的利用。图4a所示的数据变动幅度曲线从第T代开始变动幅度不大，在训练的中后期不用代代都重新确认量化参数，在第T代或者第T+1代，利用当前代对应量化前的数据以及量化后的数据确定量化误差，根据量化误差确定对数据位宽是否需要调整以及如何调整，还要根据公式(17)确定目标迭代间隔。如果确认的目标迭代间隔为55代，这就要求从第T代或第T+1之后隔55代对应的时间点作为第一预判时间点再判断是否调整数据位宽以及如何调整，并根据公式(17)确定下一目标迭代间隔，从而确定下一个第一预判时间点，直至同一周期(epoch)内所有代运算完成。在此基础上，在每个周期(epoch)之后，再对数据位宽或量化参数做适应性调整，最终使用量化

后的数据获得精度符合预期的神经网络。

[0136] 在该实现方式中，假如：根据图4a所示的权值数据变动幅度曲线图确定T取值为130(这个数值与图4a不对应，为方便描述，仅仅是假设T取值为130，不限于在假设值。)，那么训练过程中的第130代作为第二预判时间点，当前的第一预判时间点为训练过程中的第100代，在第100代，经公式(17)确定目标迭代间隔为35代。在该目标迭代间隔内，训练至第130代，到达第二预判时间点，此时就要在第130代对应的时间点确定对数据位宽是否需要调整以及如何调整，还要根据公式(17)确定目标迭代间隔。假如该情况下确定的目标迭代间隔为42代。就要从第130代起至第172代作为目标迭代间隔，目标迭代间隔为35代时确定的第一预判时间点对应的第135代处于目标迭代间隔为42代内，在第135代，可以再根据公式(17)判断是否需要调整数据位宽以及如何调整。也可以不在第135代做评估预判，直接到第172代再执行是否需要调整数据位宽的评估以及如何调整。总之，是否在第135代进行评估和预判均适于本公开所提供的技术方案。

[0137] 综上，根据数据变动幅度曲线提前预设第二预判时间点，在训练或微调的初期，无需花费人工智能处理器芯片的资源来确定目的迭代间隔，在预设好的第二预判时间点上直接根据量化误差来调整数据位宽，并利用调整好的数据位宽来量化当前代涉及的待量化数据。在训练或微调的中后期，根据公式(17)获得目标迭代间隔，从而确定对应的第一预判时间点，在每个第一预判时间点上确定是否调整数据位宽以及如何调整。这样在能够满足神经网络运算所需的浮点运算的精度的同时合理利用人工智能处理器芯片的资源，大大提高了量化时的效率。

[0138] 在一种可能的实现方式中，为了获得更准确的数据位宽的目标迭代间隔，不仅仅根据点位置参数的变化趋势值，可以同时考虑点位置参数的变化趋势值和数据位宽的变化趋势值。目标迭代间隔的确定步骤可以包括：

[0139] 在预判时间点，确定权值迭代过程中待量化数据对应点位置参数的变化趋势值、数据位宽的变化趋势值；其中，预判时间点是用于判断是否需要对数据位宽进行调整的时间点，预判时间点对应权值更新迭代完成时的时间点；

[0140] 根据点位置参数的变化趋势值和数据位宽的变化趋势值确定对应目标迭代间隔。

[0141] 在该实现方式中，可以根据式(18)来利用对应量化误差确定数据位宽的变化趋势值。

[0142] $$diff_{update2} = \delta * diff_{bit}^{2} \qquad (18)$$

[0143] 式(19)中，δ为超参数，$diff_{bit}$为量化误差；$diff_{update2}$为数据位宽的变化趋势值。$diff_{update2}$衡量量化时采用的数据位宽n的变化趋势，$diff_{update2}$越大越有可能需要更新定点的位宽，需要间隔更短的更新频率。

[0144] 在该实现方式中，点位置参数的变化趋势值仍然可根据式(15)获得，对于式(15)中的$M^{(t)}$根据公式(16)获得。$diff_{update1}$衡量点位置参数s变化趋势，由于点位置参数s的变化也变相体现在当前待量化数据中数据最大值Z_{max}的变化情况。$diff_{update1}$越大，说明数值范围变化剧烈，需要间隔更短的更新频率，即目标迭代间隔更小。

[0145] 在该实现方式中，根据式(19)确定目标迭代间隔。对于目标迭代间隔来说，同一目标迭代间隔内量化过程中采用相同的数据位宽，不同目标迭代间隔内量化过程中采用的数

据位宽可以相同，也可以不同。

[0146] $$I=\frac{\beta}{\max\left(diff_{update1},diff_{update2}\right)}-\gamma \qquad (19)$$

[0147] 式(19)中，I为目标迭代间隔。β、γ为超参数。$diff_{update1}$为点位置参数的变化趋势值。$diff_{update2}$为数据位宽的变化趋势值。

[0148] 在该实现方式中，，$diff_{update1}$是用来衡量点位置参数s的变化情况，但是由数据位宽n的变化而导致的点位置参数s的变化是要忽略掉的。因为这已经在$diff_{update2}$中体现过了数据位宽n的变化。如果在$diff_{update1}$中不做这个忽略的操作，那么根据式(19)确定的目标迭代间隔I是不准确的，造成第一预判时间点过多，在训练或微调过程中，易频繁的做数据位宽n是否更新以及如何更新的操作，从而造成人工智能处理器芯片的资源没有合理利用。

[0149] 在该实现方式中，$diff_{update1}$根据$M^{(t)}$确定。假设第t-1个预判时间点对应的数据位宽为n_1，对应的点位置参数为s_1，点位置参数随着训练迭代增加的滑动平均值为m_1。利用数据位宽n_1对待量化数据进行量化，获得量化后的定点数。根据量化前的数据和对应的量化后的数据，确定量化误差$diff_{bit}$，根据量化误差$diff_{bit}$与阈值的比较结果，将数据位宽n_1调整为n_2，数据位宽调整了$|n_1-n_2|$位，第t个预判时间点量化时使用的数据位宽为n_2。为了忽略由数据位宽的变化而导致的点位置参数的变化，在确定$M^{(t)}$时可以选出下述两种优化方式中的其中一种即可。第一种方式：如果数据位宽增加了$|n_1-n_2|$位，则$s^{(t-1)}$取值为$s_1-|n_1-n_2|$，$M^{(t-1)}$取值为$m_1-|n_1-n_2|$，将$s^{(t-1)}$、$M^{(t-1)}$代入公式(16)中，获得$M^{(t)}$，即为第t个预判时间点对应的点位置参数随着训练迭代增加的滑动平均值。如果数据位宽减少了$|n_1-n_2|$位，则$s^{(t-1)}$取值为$s_1+|n_1-n_2|$，$M^{(t-1)}$取值为$m_1+|n_1-n_2|$，将$s^{(t-1)}$、$M^{(t-1)}$代入公式(16)中，获得$M^{(t)}$，即为第t个预判时间点对应的点位置参数随着训练迭代增加的滑动平均值。第二种方式：不管数据位宽是增加了$|n_1-n_2|$位还是减少了$|n_1-n_2|$，$s^{(t-1)}$取值为s_1，$M^{(t-1)}$取值为m_1，将$s^{(t-1)}$、$M^{(t-1)}$代入公式(16)中，获得$M^{(t)}$。在数据位宽增加$|n_1-n_2|$位时，将$M^{(t)}$减去$|n_1-n_2|$，在数据位宽减少$|n_1-n_2|$位时，将$M^{(t)}$加上$|n_1-n_2|$，结果作为第t个预判时间点对应的点位置参数随着训练迭代增加的滑动平均值。这两种方式是等价的，均可以忽略由数据位宽的变化而导致的点位置参数的变化，获得更为精准的目标迭代间隔，从而提高人工智能处理器芯片的资源利用率。

[0150] 在实际应用中，数据位宽n和点位置参数s对量化影响很大，量化参数中的缩放系数f以及偏移量0对量化影响不大。所以，不管数据位宽n是否发生变化、点位置参数s可变的情况下，确定点位置参数s的目标迭代间隔也是一件非常有意义的事情。

[0151] 在一种可能的实现方式中，确定目标迭代间隔的过程可以包括以下步骤：

[0152] 在预判时间点，确定权值迭代过程中涉及的待量化数据对应点位置参数的变化趋势值；其中，预判时间点是用于判断是否需要对量化参数进行调整的时间点，预判时间点对应权值更新迭代完成时的时间点；

[0153] 根据点位置参数的变化趋势值确定对应目标迭代间隔。

[0154] 在该实现方式中，量化参数优选为点位置参数。

[0155] 在步骤S12中，根据所述量化后的数据继续执行winograd卷积处理，得到量化后的winograd卷积结果。

[0156] 在步骤S13中，对所述量化后的winograd卷积结果执行反量化处理，得到winograd卷积结果。

[0157] 在一种可能的实现方式中，所述winograd卷积处理过程，包括：

[0158] 将待运算数据的winograd正变换拆解为求和运算，并进行计算得到所述待运算数据中每个数据的winograd正变换结果；

[0159] 执行所述待运算数据中对应数据的winograd正变换结果之间的对位乘操作，得到对位乘结果；

[0160] 将对所述对位乘结果的winograd逆变换拆解为求和运算，得到所述winograd卷积结果，

[0161] 其中，所述待量化数据可以为所述待运算数据、所述待运算数据的winograd正变换结果和所述对位乘结果中的任一种。

[0162] 在一种可能的实现方式中，所述待运算数据包括输入神经元、权值和梯度中的至少一种。

[0163] 在本公开实施例中所提及的待量化数据、待运算数据等数据可以是实际数据处理过程中所能够出现的数据，可以是与图像数据、视频数据、音频数据、文本数据等数据相对应的数据。例如，利用本公开所提供的方法进行图像处理、视频处理、音频处理等场景的应用。以待运算数据为图像数据为例，待运算数据可以表示为NHWC(batch,height,width,channels)的形式，N表示图像的数量，HW分别表示在高度和宽度方向的像素个数，C可以表示通道数，例如，C可以表示RGB(Red,Ggreen,Blue)三个通道，上述表示方式仅为本公开的一个示例，本公开不限于此。需要说明的是，上述方法可以应用于任何可以量化、进行winograd卷积运算处理的数据，本领域技术人员可以根据实际需要对数据的类型及其对应的应用场景进行设置，本公开对此不作限制。

[0164] 举例来说，可以对待量化数据进行量化，从而加快winograd卷积的处理速度。在一些实施例中，待量化数据可以为32位的浮点数。备选地，待量化的数据也可以为其他位数的浮点数，或者其他的数据类型。

[0165] 在一种可能的实现方式中，根据确定出的所述一对截断阈值对所述待量化数据进行量化处理，得到量化后的数据，包括以下任一操作：

[0166] 在将待运算数据的winograd正变换拆解为求和运算之前，将所述待运算数据作为待量化数据进行量化处理；

[0167] 在进行对位乘操作之前，将待运算数据中每个数据的winograd正变换结果作为待量化数据进行量化处理；

[0168] 在进行winograd逆变换拆解为求和运算之前，将所述对位乘结果作为待量化数据进行量化处理。

[0169] 示例性的，待量化数据为待运算数据，则winograd卷积过程可以为：

[0170] 采用确定的一对截断阈值量化待运算数据，得到量化后的待运算数据；将量化后的待运算数据的winograd正变换拆解为求和运算，并进行计算得到量化后的待运算数据的winograd正变换结果；执行量化后的待运算数据的winograd正变换结果的对位乘操作，得到对位乘结果；将对所述对位乘结果的winograd逆变换拆解为求和运算，得到所述量化后的winograd卷积结果，对该量化后的winograd卷积结果进行反量化处理，得到winograd卷

积结果。

[0171] 示例性的，待量化数据为待运算数据的winograd正变换结果，则winograd卷积过程可以为：

[0172] 将待运算数据的winograd正变换拆解为求和运算，并进行计算得到待运算数据的winograd正变换结果；采用确定的一对截断阈值量化待运算数据的winograd正变换结果，得到量化后的待运算数据的winograd正变换结果；执行量化后的待运算数据的winograd正变换结果的对位乘操作，得到对位乘结果；将对对位乘结果的winograd逆变换拆解为求和运算，得到所述量化后的winograd卷积结果，对该量化后的winograd卷积结果进行反量化处理，得到winograd卷积结果。

[0173] 示例性的，待量化数据为对位乘结果，则winograd卷积过程可以为：

[0174] 将待运算数据的winograd正变换拆解为求和运算，并进行计算得到待运算数据的winograd正变换结果；执行待运算数据的winograd正变换结果的对位乘操作，得到对位乘结果；采用确定的一对截断阈值量化对位乘结果，得到量化后的对位乘结果；将对量化后的对位乘结果的winograd逆变换拆解为求和运算，得到量化后的winograd卷积结果。对该量化后的winograd卷积结果进行反量化处理，得到winograd卷积结果。

[0175] 在一种可能的实现方式中，所述将待运算数据的winograd正变换拆解为求和运算，并进行计算得到所述待运算数据中每个数据的winograd正变换结果，包括：

[0176] 将所述待运算数据中的每个数据分别拆解为多个第一子张量，对所述待运算数据中的每个数据的多个第一子张量进行winograd正变换并求和得到所述待运算数据中每个数据的winograd正变换结果，

[0177] 其中，每个数据拆分出的多个第一子张量的个数与对应的数据中不为0元素的个数相同，每个第一子张量中有一个元素与对应数据中的对应位置的元素相同、其他元素均为0。

[0178] 举例来说，假设输入神经元表示为：

[0179] $d_{4\times4}=\begin{bmatrix} d_{00} & d_{01} & d_{02} & d_{03} \\ d_{10} & d_{11} & d_{12} & d_{13} \\ d_{20} & d_{21} & d_{22} & d_{23} \\ d_{30} & d_{31} & d_{32} & d_{33} \end{bmatrix}$，输入神经元为4×4的矩阵，包括16个元素，因此，可以将待运算数据拆解为16个第一子张量。

[0180] 那么，按照本公开的拆解方式，16个第一子张量分别为：

[0181] $d_{00}=\begin{bmatrix} d_{00} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$，$d_{01}=\begin{bmatrix} 0 & d_{01} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$，$d_{02}=\begin{bmatrix} 0 & 0 & d_{02} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$，$d_{03}=\begin{bmatrix} 0 & 0 & 0 & d_{03} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$……，

$d_{33}=\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & d_{33} \end{bmatrix}$

[0182] 每个第一子张量中有一个元素与所述待运算数据中的对应位置的元素相同、其他元素均为0是指：以第一子张量d_{00}为例，在第一行第一列位置的元素与输入神经元在第一行第一列的位置的元素相同，其他元素都为0，其他第一子张量也有相同的属性。

[0183] 需要说明的是，以上拆解方式仅仅是本公开的一些示例，不以任何方式限制本公开，例如，如果待运算数据中具有值为0的元素，拆解得到的第一子张量的数量可以少于待运算数据的元素的个数，例如，多个第一子张量的个数与所述待运算数据的不为0的元素的个数相同。

[0184] 在一种可能的实现方式中，对所述待运算数据中的每个数据的多个第一子张量进行winograd正变换并求和得到所述待运算数据中每个数据的winograd正变换结果，可以包括：

[0185] 获取第一子张量对应的第一元子张量的winograd正变换结果；其中，第一子张量对应的第一元子张量为：在第一元子张量中第一位置的元素的值为1，其中，第一位置在第一元子张量中所处的位置与第一子张量中的非0元素所处的位置相同；

[0186] 将第一子张量中不为0的元素值作为系数乘以对应的第一元子张量的winograd正变换结果，得到第一子张量的winograd正变换结果；

[0187] 将多个第一子张量的winograd正变换结果相加得到所述待运算数据的winograd正变换结果。

[0188] 仍然以第一子张量d_{00}为例，d_{00}对应的第一元子张量可以为$\begin{bmatrix}1&0&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&0\end{bmatrix}$，也就是说，第一元子张量是将第一子张量中的非0元素值提取出来，非0元素的值可以作为第一元子张量的系数。

[0189] 其中，第一子张量对应的第一元子张量的winograd正变换结果可以是通过以下过程预先得到的：对于每一个第一子张量，将该第一子张量对应的第一元子张量左边乘以正变换左乘矩阵、右边乘以正变换右乘矩阵得到第一元子张量的winograd正变换结果。

[0190] 对于不同尺寸的矩阵，对应的第一元子张量的形式是确定的，对应的正变换左乘矩阵和正变换右乘矩阵也是确定的。

[0191] 因此，可以预先计算出第一元子张量的winograd正变换结果，具体过程如上所述。举例来说，仍然以d_{00}为例，其对应的第一元子张量的winograd正变换结果为：

[0192] $$\begin{bmatrix}1&0&-1&0\\0&1&1&0\\0&-1&1&0\\0&1&0&-1\end{bmatrix}\begin{bmatrix}1&0&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&0\end{bmatrix}\begin{bmatrix}1&0&0&0\\0&1&-1&1\\-1&1&1&0\\0&0&0&-1\end{bmatrix}=\begin{bmatrix}1&0&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&0\end{bmatrix}。$$

[0193] 再比如，以d_{01}为例，其对应的第一元子张量的winograd正变换结果为：

[0194] $$\begin{bmatrix}1&0&-1&0\\0&1&1&0\\0&-1&1&0\\0&1&0&-1\end{bmatrix}\begin{bmatrix}0&1&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&0\end{bmatrix}\begin{bmatrix}1&0&0&0\\0&1&-1&1\\-1&1&1&0\\0&0&0&-1\end{bmatrix}=\begin{bmatrix}0&1&-1&1\\0&0&0&0\\0&0&0&0\\0&0&0&0\end{bmatrix}。$$

[0195] 由于正变换左乘矩阵和正变换右乘矩阵的元素值都是0、±1，第一元子张量的元素值为0或1，第一元子张量的winograd正变换结果中的元素也是0、±1。因此，可以将矩阵乘操作拆解为加法操作。

[0196] 计算第一元子张量的winograd正变换结果的过程涉及较多的乘法运算，通过本公开的方式，可以将预先计算好的各种规模的第一元子张量的winograd正变换结果保存在运算装置中，这样，在实际的运算过程中，可以直接获取，而不需要重复运算，从而缩短计算时

间、节约计算资源。

[0197] 在获得第一子张量对应的第一元子张量的winograd正变换结果，可以将第一子张量中不为0的元素值乘以对应的第一元子张量的winograd正变换结果，就可以得到第一子张量的winograd正变换结果。举例来说，仍然以d_{00}为例，其对应的winograd正变换结果为：$d_{00}\begin{bmatrix}1&0&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&0\end{bmatrix}$。

[0198] 再比如，以d_{01}为例，d_{01}的winograd正变换结果为$d_{01}\begin{bmatrix}0&1&-1&1\\0&0&0&0\\0&0&0&0\\0&0&0&0\end{bmatrix}$。

[0199] 通过以上过程计算得到所有第一子张量的winograd正变换结果，将多个第一子张量的winograd正变换结果相加，即可得到所述待运算数据的winograd正变换结果。

[0200] $$B^{T}d_{4\times4}B=d_{00}\begin{bmatrix}1&0&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&0\end{bmatrix}+d_{01}\begin{bmatrix}0&1&-1&1\\0&0&0&0\\0&0&0&0\\0&0&0&0\end{bmatrix}+\cdots+d_{33}\begin{bmatrix}0&0&0&0\\0&0&0&0\\0&0&0&0\\0&0&0&1\end{bmatrix}\quad(20)$$

[0201] $$G_{4\times3}g_{3\times3}G_{3\times4}^{T}=g_{00}\begin{bmatrix}1&1&1&0\\1&1&1&0\\1&1&1&0\\0&0&0&0\end{bmatrix}+g_{01}\begin{bmatrix}0&1&-1&0\\0&1&-1&0\\0&1&-1&0\\0&0&0&0\end{bmatrix}+\cdots+g_{22}\begin{bmatrix}0&0&0&0\\0&1&1&1\\0&1&1&1\\0&1&1&1\end{bmatrix}\quad(21)$$

[0202] 由于转换得到的第一元子张量的winograd正变换结果中的元素也是0、±1，因此，上述等式(20)、(21)右侧仅涉及求和运算。

[0203] 根据本公开上述实施方式可知，通过将待运算数据进行拆解得到多个第一子张量，根据预先计算得到的第一子张量对应的第一元子张量的winograd正变换结果以及第一子张量的非0元素值即可进行求和运算得到待运算数据的winograd正变换结果。

[0204] 在采用上文提到的拆解为求和运算得到输入神经元的winograd正变换结果后，可以采用计算权值的winograd正变换结果，其中权值的winograd正变换结果的计算方式可以采用传统的矩阵乘法计算，也可以参照上文提到的拆解为求和运算进行计算得到winograd正变换结果。

[0205] 在得到待运算数据(输入神经元、权值、梯度)的winograd正变换结果后，可以继续执行待运算数据的winograd正变换结果的对位乘操作，得到对位乘结果。其中，对位乘可以是指对两个张量对应位置的数据相乘得到的数据作为对位乘结果中相应位置的值。

[0206] 假设输入神经元的winograd正变换结果$B^{T}d_{4\times4}B$可以表示为：$D_{4\times4}=\begin{bmatrix}D_{00}&D_{01}&D_{02}&D_{03}\\D_{10}&D_{11}&D_{12}&D_{13}\\D_{20}&D_{21}&D_{22}&D_{23}\\D_{30}&D_{31}&D_{32}&D_{33}\end{bmatrix}$，

[0207] 权值的winograd正变换结果$G_{4\times3}g_{3\times3}G_{3\times4}^{T}$可以表示为：$G_{4\times4}=\begin{bmatrix}G_{00}&G_{01}&G_{02}&G_{03}\\G_{10}&G_{11}&G_{12}&G_{13}\\G_{20}&G_{21}&G_{22}&G_{23}\\G_{30}&G_{31}&G_{32}&G_{33}\end{bmatrix}$，

[0208] 那么对位乘结果可以为：

[0209] $$G_{4\times4}\odot D_{4\times4}=\begin{bmatrix} D_{00}\times G_{00} & D_{01}\times G_{01} & D_{02}\times G_{02} & D_{03}\times G_{03} \\ D_{10}\times G_{10} & D_{11}\times G_{11} & D_{12}\times G_{12} & D_{13}\times G_{13} \\ D_{20}\times G_{20} & D_{21}\times G_{21} & D_{22}\times G_{22} & D_{23}\times G_{23} \\ D_{30}\times G_{30} & D_{31}\times G_{31} & D_{32}\times G_{32} & D_{33}\times G_{33} \end{bmatrix}。$$

[0210] 待运算数据的winograd卷积结果可以表示为$S_{4\times4}=A^{T}(G_{4\times4}\odot D_{4\times4})A$，本公开的从功能处理单元可以将$A^{T}(G_{4\times4}\odot D_{4\times4})A$拆解为求和运算，并进行计算得到所述待运算数据的winograd卷积结果，从而可以进一步节约计算时间、减少能耗。

[0211] 在一种可能的实现方式中，所述将对所述对位乘结果的winograd逆变换拆解为求和运算，得到所述winograd卷积结果，可以包括：

[0212] 将所述对位乘结果拆解为多个第二子张量，对所述多个第二子张量进行winograd逆变换并求和，得到所述待运算数据的winograd卷积结果；

[0213] 其中，所述多个第二子张量的个数与所述对位乘结果中不为0元素的个数相同，所述多个第二子张量中的每个第二子张量中有一个元素与所述对位乘结果中的对应位置的元素相同、其他元素均为0。

[0214] 假设对位乘结果为：

[0215] $$C_{4\times4}=G_{4\times4}\odot D_{4\times4}=\begin{bmatrix} D_{00}\times G_{00} & D_{01}\times G_{01} & D_{02}\times G_{02} & D_{03}\times G_{03} \\ D_{10}\times G_{10} & D_{11}\times G_{11} & D_{12}\times G_{12} & D_{13}\times G_{13} \\ D_{20}\times G_{20} & D_{21}\times G_{21} & D_{22}\times G_{22} & D_{23}\times G_{23} \\ D_{30}\times G_{30} & D_{31}\times G_{31} & D_{32}\times G_{32} & D_{33}\times G_{33} \end{bmatrix}=\begin{bmatrix} C_{00} & C_{01} & C_{02} & C_{03} \\ C_{10} & C_{11} & C_{12} & C_{13} \\ C_{20} & C_{21} & C_{22} & C_{23} \\ C_{30} & C_{31} & C_{32} & C_{33} \end{bmatrix}$$

[0216] 将对位乘结果拆解为多个第二子张量，例如可以拆解为16个，16个第二子张量分别为：

[0217] $$C_{00}=\begin{bmatrix} C_{00} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},\quad C_{01}=\begin{bmatrix} 0 & C_{01} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},\quad C_{02}=\begin{bmatrix} 0 & 0 & C_{02} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},\quad C_{03}=\begin{bmatrix} 0 & 0 & 0 & C_{03} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}\ldots\ldots,$$

$$C_{33}=\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & C_{33} \end{bmatrix}。$$

[0218] 在拆解完后，可以对所述多个第二子张量进行winograd逆变换并求和得到所述待运算数据的winograd卷积结果。

[0219] 在一种可能的实现方式中，对所述多个第二子张量进行winograd逆变换并求和得到所述待运算数据的winograd卷积结果，可以包括以下过程：

[0220] 获取第二子张量对应的第二元子张量的winograd逆变换结果；其中，第二子张量对应的第二元子张量为：在第二元子张量中第二位置的元素的值为1，其中，第二位置在第二元子张量中所处的位置与第二子张量中的非0元素所处的位置相同；

[0221] 将第二子张量中不为0的元素值作为系数乘以对应的第二元子张量的winograd逆变换结果，得到第二子张量的winograd逆变换结果；

[0222] 将多个第二子张量的winograd逆变换结果相加得到所述待运算数据的winograd卷积结果。

[0223] 第二子张量对应的第二元子张量确定的方式和上文中第一元子张量确定的方式相同，不再赘述。其中，第二元子张量的winograd逆变换结果是通过以下过程预先得到的：

对于每一个第二子张量，将该第二子张量对应的第二元子张量左边乘以逆变换左乘矩阵、右边乘以逆变换右乘矩阵得到第二元子张量的winograd逆变换结果。

[0224] 对于不同尺寸的矩阵，对应的第二元子张量的形式是确定的，对应的逆变换左乘矩阵和逆变换右乘矩阵也是确定的。因此，可以预先计算出第二元子张量的winograd逆变换结果，具体过程如上所述。

[0225] 对于本文上述列举的示例，逆变换左乘矩阵为2×4的矩阵，例如可以为：$\begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & -\frac{1}{2} & -1 \end{bmatrix}$，

[0226] 逆变换右乘矩阵为4×2的矩阵，例如可以为：$\begin{bmatrix} 1 & 0 \\ \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \\ 0 & -1 \end{bmatrix}$。

[0227] 逆变换矩阵的维度可以根据输入神经元的维度以及权值的维度和卷积步长确定，上文仅仅是一个示例，不以任何方式限制本公开。

[0228] 逆变换矩阵由0，$\pm\frac{1}{2}$，±1构成，因此逆变换的矩阵乘操作可以拆解为加法和移位操作实现。将逆变换矩阵乘以第二元子张量即可得到第二元子张量的winograd逆变换结果，第二元子张量的winograd逆变换结果内的元素值由0，$\pm\frac{1}{4}$，$\pm\frac{1}{2}$，±1等构成，分数可以通过简单的移位操作计算，相比于乘法操作仍然可以节省计算时间。

[0229] 对于“将第二子张量中不为0的元素值作为系数乘以对应的第二元子张量的winograd逆变换结果，得到第二子张量的winograd逆变换结果；将多个第二子张量的winograd逆变换结果相加得到所述待运算数据的winograd卷积结果”的具体过程可以参照上文，只不过第二元子张量的winograd逆变换结果不完全由0、±1，但分数可以通过简单的移位操作计算，相比于乘法操作，本公开将普通的逆变换过程拆解后仍然可以实现节约计算时间、减少能耗的效果。

[0230] 根据本公开上述实施方式可知，通过将对位乘结果进行拆解得到多个第二子张量，根据预先计算得到的第二子张量对应的第二元子张量的winograd逆变换结果以及第二子张量的非0元素值即可进行求和运算得到待运算数据的winograd卷积结果。

[0231] 根据本公开的数据处理方法，根据确定出的量化参数对待量化数据进行量化处理，得到量化后的数据，其中，量化参数是根据对应的待量化数据的统计结果和数据位宽确定的；根据量化后的数据继续执行winograd卷积处理，得到量化后的winograd卷积结果；对量化后的winograd卷积结果执行反量化处理，得到winograd卷积结果。可以提高量化的精度的同时，节约winograd卷积的运算时间，减少能耗。

[0232] 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。

[0233] 进一步需要说明的是，虽然图1的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，图1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。

[0234] 图5示出根据本公开实施例的数据处理装置的框图。如图5所示，该装置包括：数据量化模块41、卷积处理模块42和反量化处理模块43。数据量化模块41，根据确定出的量化参数对待量化数据进行量化处理，得到量化后的数据，其中，所述量化参数是根据对应的待量化数据的统计结果和数据位宽确定的。卷积处理模块42，根据所述量化后的数据继续执行winograd卷积处理，得到量化后的winograd卷积结果。反量化处理模块43，对所述量化后的winograd卷积结果执行反量化处理，得到winograd卷积结果。

[0235] 在一种可能的实现方式中，所述winograd卷积处理过程，包括：将待运算数据的winograd正变换拆解为求和运算，并进行计算得到所述待运算数据中每个数据的winograd正变换结果；执行所述待运算数据中对应数据的winograd正变换结果之间的对位乘操作，得到对位乘结果；将对所述对位乘结果的winograd逆变换拆解为求和运算，得到所述winograd卷积结果，其中，所述待量化数据为所述待运算数据、所述待运算数据的winograd正变换结果和所述对位乘结果中的任一种。

[0236] 在一种可能的实现方式中，根据确定出的量化参数对待量化数据进行量化处理，得到量化后的数据，包括以下任一操作：在将待运算数据的winograd正变换拆解为求和运算之前，将所述待运算数据作为待量化数据进行量化处理；在进行对位乘操作之前，将待运算数据中每个数据的winograd正变换结果作为待量化数据进行量化处理；在进行winograd逆变换拆解为求和运算之前，将所述对位乘结果作为待量化数据进行量化处理。

[0237] 在一种可能的实现方式中，所述将待运算数据的winograd正变换拆解为求和运算，并进行计算得到所述待运算数据中每个数据的winograd正变换结果，包括：将所述待运算数据中的每个数据分别拆解为多个第一子张量，对所述待运算数据中的每个数据的多个第一子张量进行winograd正变换并求和得到所述待运算数据中每个数据的winograd正变换结果，其中，每个数据拆分出的多个第一子张量的个数与对应的数据中不为0元素的个数相同，每个第一子张量中有一个元素与对应数据中的对应位置的元素相同、其他元素均为0。

[0238] 在一种可能的实现方式中，所述将对所述对位乘结果的winograd逆变换拆解为求和运算，得到所述winograd卷积结果，包括：将所述对位乘结果拆解为多个第二子张量，对所述多个第二子张量进行winograd逆变换并求和，得到所述待运算数据的winograd卷积结果；其中，所述多个第二子张量的个数与所述对位乘结果中不为0元素的个数相同，所述多个第二子张量中的每个第二子张量中有一个元素与所述对位乘结果中的对应位置的元素相同、其他元素均为0。

[0239] 在一种可能的实现方式中，所述统计结果包括以下任一种：每种待量化数据中的绝对值最大值、每种待量化数据中的最大值和最小值的距离的二分之一，所述量化参数包括点位置参数、缩放系数和偏移量中的一种或多种，其中，所述绝对值最大值是每种待量化

数据中的最大值或最小值的绝对值。

[0240] 在一种可能的实现方式中,所述缩放系数是根据所述点位置参数、所述统计结果、所述数据位宽确定的。

[0241] 在一种可能的实现方式中,所述偏移量是根据每种待量化数据的统计结果确定的。

[0242] 在一种可能的实现方式中,所述点位置参数是根据所述统计结果和所述数据位宽确定的。

[0243] 在一种可能的实现方式中,所述装置还包括:位宽调整模块,根据所述数据位宽对应的量化误差,对所述数据位宽进行调整,以利用调整后的数据位宽确定量化参数,其中,所述量化误差是根据对应层中量化后的数据与对应的量化前的数据确定的。

[0244] 在一种可能的实现方式中,所述量化前的数据是在目标迭代间隔内的权值更新迭代过程中涉及的待量化数据;其中,所述目标迭代间隔包括至少一次权值更新迭代,且同一目标迭代间隔内量化过程中采用相同的所述数据位宽。

[0245] 在一种可能的实现方式中,所述待运算数据包括输入神经元、权值和梯度中的至少一种。

[0246] 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。

[0247] 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。

[0248] 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(Resistive 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。

[0249] 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。

[0250] 在一种可能的实现方式中，还公开了一种人工智能芯片，其包括了上述数据处理装置。

[0251] 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。

[0252] 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；

[0253] 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)。

[0254] 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。

[0255] 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。

[0256] 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PCIE3.0X 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)。

[0257] 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。

[0258] 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。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。

[0259] 本公开实施例还提出一种计算机可读存储介质,其上存储有计算机程序指令,所述计算机程序指令被处理器执行时实现上述方法。计算机可读存储介质可以是非易失性计算机可读存储介质。

[0260] 本公开实施例还提出一种电子设备,包括:处理器;用于存储处理器可执行指令的存储器;其中,所述处理器被配置为调用所述存储器存储的指令,以执行上述方法。

[0261] 图7示出根据本公开实施例的一种电子设备800的框图。例如,电子设备800可以是移动电话,计算机,数字广播终端,消息收发设备,游戏控制台,平板设备,医疗设备,健身设备,个人数字助理等终端。

[0262] 参照图7,电子设备800可以包括以下一个或多个组件:处理组件802,存储器804,电源组件806,多媒体组件808,音频组件810,输入/输出(I/O)的接口812,传感器组件814,以及通信组件816。

[0263] 处理组件802通常控制电子设备800的整体操作,诸如与显示,电话呼叫,数据通信,相机操作和记录操作相关联的操作。处理组件802可以包括一个或多个处理器820来执行指令,以完成上述的方法的全部或部分步骤。此外,处理组件802可以包括一个或多个模块,便于处理组件802和其他组件之间的交互。例如,处理组件802可以包括多媒体模块,以方便多媒体组件808和处理组件802之间的交互。

[0264] 存储器804被配置为存储各种类型的数据以支持在电子设备800的操作。这些数据的示例包括用于在电子设备800上操作的任何应用程序或方法的指令,联系人数据,电话簿数据,消息,图片,视频等。存储器804可以由任何类型的易失性或非易失性存储设备或者它们的组合实现,如静态随机存取存储器(SRAM),电可擦除可编程只读存储器(EEPROM),可擦除可编程只读存储器(EPROM),可编程只读存储器(PROM),只读存储器(ROM),磁存储器,快闪存储器,磁盘或光盘。

[0265] 电源组件806为电子设备800的各种组件提供电力。电源组件806可以包括电源管理系统,一个或多个电源,及其他与为电子设备800生成、管理和分配电力相关联的组件。

[0266] 多媒体组件808包括在所述电子设备800和用户之间的提供一个输出接口的屏幕。在一些实施例中,屏幕可以包括液晶显示器(LCD)和触摸面板(TP)。如果屏幕包括触摸面板,屏幕可以被实现为触摸屏,以接收来自用户的输入信号。触摸面板包括一个或多个触摸传感器以感测触摸、滑动和触摸面板上的手势。所述触摸传感器可以不仅感测触摸或滑动动作的边界,而且还检测与所述触摸或滑动操作相关的持续时间和压力。在一些实施例中,多媒体组件808包括一个前置摄像头和/或后置摄像头。当电子设备800处于操作模式,如拍摄模式或视频模式时,前置摄像头和/或后置摄像头可以接收外部的多媒体数据。每个前置摄像头和后置摄像头可以是一个固定的光学透镜系统或具有焦距和光学变焦能力。

[0267] 音频组件810被配置为输出和/或输入音频信号。例如,音频组件810包括一个麦克风(MIC),当电子设备800处于操作模式,如呼叫模式、记录模式和语音识别模式时,麦克风被配置为接收外部音频信号。所接收的音频信号可以被进一步存储在存储器804或经由通信组件816发送。在一些实施例中,音频组件810还包括一个扬声器,用于输出音频信号。

[0268] I/O接口812为处理组件802和外围接口模块之间提供接口，上述外围接口模块可以是键盘，点击轮，按钮等。这些按钮可包括但不限于：主页按钮、音量按钮、启动按钮和锁定按钮。

[0269] 传感器组件814包括一个或多个传感器，用于为电子设备800提供各个方面的状态评估。例如，传感器组件814可以检测到电子设备800的打开/关闭状态，组件的相对定位，例如所述组件为电子设备800的显示器和小键盘，传感器组件814还可以检测电子设备800或电子设备800一个组件的位置改变，用户与电子设备800接触的存在或不存在，电子设备800方位或加速/减速和电子设备800的温度变化。传感器组件814可以包括接近传感器，被配置用来在没有任何的物理接触时检测附近物体的存在。传感器组件814还可以包括光传感器，如CMOS或CCD图像传感器，用于在成像应用中使用。在一些实施例中，该传感器组件814还可以包括加速度传感器，陀螺仪传感器，磁传感器，压力传感器或温度传感器。

[0270] 通信组件816被配置为便于电子设备800和其他设备之间有线或无线方式的通信。电子设备800可以接入基于通信标准的无线网络，如WiFi，2G或3G，或它们的组合。在一个示例性实施例中，通信组件816经由广播信道接收来自外部广播管理系统的广播信号或广播相关信息。在一个示例性实施例中，所述通信组件816还包括近场通信(NFC)模块，以促进短程通信。例如，在NFC模块可基于射频识别(RFID)技术，红外数据协会(IrDA)技术，超宽带(UWB)技术，蓝牙(BT)技术和其他技术来实现。

[0271] 在示例性实施例中，电子设备800可以被一个或多个应用专用集成电路(ASIC)、数字信号处理器(DSP)、数字信号处理设备(DSPD)、可编程逻辑器件(PLD)、现场可编程门阵列(FPGA)、控制器、微控制器、微处理器或其他电子元件实现，用于执行上述方法。

[0272] 在示例性实施例中，还提供了一种非易失性计算机可读存储介质，例如包括计算机程序指令的存储器804，上述计算机程序指令可由电子设备800的处理器820执行以完成上述方法。

[0273] 图8示出根据本公开实施例的一种电子设备1900的框图。例如，电子设备1900可以被提供为一服务器。参照图8，电子设备1900包括处理组件1922，其进一步包括一个或多个处理器，以及由存储器1932所代表的存储器资源，用于存储可由处理组件1922的执行的指令，例如应用程序。存储器1932中存储的应用程序可以包括一个或一个以上的每一个对应于一组指令的模块。此外，处理组件1922被配置为执行指令，以执行上述方法。

[0274] 电子设备1900还可以包括一个电源组件1926被配置为执行电子设备1900的电源管理，一个有线或无线网络接口1950被配置为将电子设备1900连接到网络，和一个输入输出(I/O)接口1958。电子设备1900可以操作基于存储在存储器1932的操作系统，例如Windows ServerTM，Mac OS XTM，UnixTM，LinuxTM，FreeBSDTM或类似。

[0275] 在示例性实施例中，还提供了一种非易失性计算机可读存储介质，例如包括计算机程序指令的存储器1932，上述计算机程序指令可由电子设备1900的处理组件1922执行以完成上述方法。

[0276] 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。

[0277] 依据以下条款可更好地理解前述内容：

[0278] 条款A1.一种数据处理方法，所述方法包括：

[0279] 根据确定出的量化参数对待量化数据进行量化处理，得到量化后的数据，其中，所述量化参数是根据对应的待量化数据的统计结果和数据位宽确定的；

[0280] 根据所述量化后的数据继续执行winograd卷积处理，得到量化后的winograd卷积结果；

[0281] 对所述量化后的winograd卷积结果执行反量化处理，得到winograd卷积结果。

[0282] 条款A2.根据条款A1所述的方法，所述winograd卷积处理过程，包括：

[0283] 将待运算数据的winograd正变换拆解为求和运算，并进行计算得到所述待运算数据中每个数据的winograd正变换结果；

[0284] 执行所述待运算数据中对应数据的winograd正变换结果之间的对位乘操作，得到对位乘结果；

[0285] 将对所述对位乘结果的winograd逆变换拆解为求和运算，得到所述winograd卷积结果，

[0286] 其中，所述待量化数据为所述待运算数据、所述待运算数据的winograd正变换结果和所述对位乘结果中的任一种。

[0287] 条款A3.根据条款A2所述的方法，根据确定出的量化参数对待量化数据进行量化处理，得到量化后的数据，包括以下任一操作：

[0288] 在将待运算数据的winograd正变换拆解为求和运算之前，将所述待运算数据作为待量化数据进行量化处理；

[0289] 在进行对位乘操作之前，将待运算数据中每个数据的winograd正变换结果作为待量化数据进行量化处理；

[0290] 在进行winograd逆变换拆解为求和运算之前，将所述对位乘结果作为待量化数据进行量化处理。

[0291] 条款A4.根据条款A2所述的方法，所述将待运算数据的winograd正变换拆解为求和运算，并进行计算得到所述待运算数据中每个数据的winograd正变换结果，包括：

[0292] 将所述待运算数据中的每个数据分别拆解为多个第一子张量，对所述待运算数据中的每个数据的多个第一子张量进行winograd正变换并求和得到所述待运算数据中每个数据的winograd正变换结果，

[0293] 其中，每个数据拆分出的多个第一子张量的个数与对应的数据中不为0元素的个数相同，每个第一子张量中有一个元素与对应数据中的对应位置的元素相同、其他元素均为0。

[0294] 条款A5.根据条款A2所述的方法，所述将对所述对位乘结果的winograd逆变换拆解为求和运算，得到所述winograd卷积结果，包括：

[0295] 将所述对位乘结果拆解为多个第二子张量，对所述多个第二子张量进行winograd逆变换并求和，得到所述待运算数据的winograd卷积结果；

[0296] 其中，所述多个第二子张量的个数与所述对位乘结果中不为0元素的个数相同，所述多个第二子张量中的每个第二子张量中有一个元素与所述对位乘结果中的对应位置的元素相同、其他元素均为0。

[0297] 条款A6.根据条款A1所述的方法，所述统计结果包括以下任一种：每种待量化数据中的绝对值最大值、每种待量化数据中的最大值和最小值的距离的二分之一，

[0298] 所述量化参数包括点位置参数、缩放系数和偏移量中的一种或多种，

[0299] 其中，所述绝对值最大值是每种待量化数据中的最大值或最小值的绝对值。

[0300] 条款A7.根据条款A6所述的方法，所述缩放系数是根据所述点位置参数、所述统计结果、所述数据位宽确定的。

[0301] 条款A8.根据条款A6所述的方法，所述偏移量是根据每种待量化数据的统计结果确定的。

[0302] 条款A9.根据条款A6所述的方法，所述点位置参数是根据所述统计结果和所述数据位宽确定的。

[0303] 条款A10.根据条款A6所述的方法，所述方法还包括：

[0304] 根据所述数据位宽对应的量化误差，对所述数据位宽进行调整，以利用调整后的数据位宽确定量化参数，

[0305] 其中，所述量化误差是根据对应层中量化后的数据与对应的量化前的数据确定的。

[0306] 条款A11.根据条款A10所述的方法，所述量化前的数据是在目标迭代间隔内的权值更新迭代过程中涉及的待量化数据；

[0307] 其中，所述目标迭代间隔包括至少一次权值更新迭代，且同一目标迭代间隔内量化过程中采用相同的所述数据位宽。

[0308] 条款A12.根据条款A2所述的方法，所述待运算数据包括输入神经元、权值和梯度中的至少一种。

[0309] 条款A13.一种数据处理装置，所述装置包括：

[0310] 数据量化模块，根据确定出的量化参数对待量化数据进行量化处理，得到量化后的数据，其中，所述量化参数是根据对应的待量化数据的统计结果和数据位宽确定的；

[0311] 卷积处理模块，根据所述量化后的数据继续执行winograd卷积处理，得到量化后的winograd卷积结果；

[0312] 反量化处理模块，对所述量化后的winograd卷积结果执行反量化处理，得到winograd卷积结果。

[0313] 条款A14.根据条款A13所述的装置，所述winograd卷积处理过程，包括：

[0314] 将待运算数据的winograd正变换拆解为求和运算，并进行计算得到所述待运算数据中每个数据的winograd正变换结果；

[0315] 执行所述待运算数据中对应数据的winograd正变换结果之间的对位乘操作，得到对位乘结果；

[0316] 将对所述对位乘结果的winograd逆变换拆解为求和运算，得到所述winograd卷积结果，

[0317] 其中，所述待量化数据为所述待运算数据、所述待运算数据的winograd正变换结果和所述对位乘结果中的任一种。

[0318] 条款A15.根据条款A14所述的装置，根据确定出的量化参数对待量化数据进行量化处理，得到量化后的数据，包括以下任一操作：

[0319] 在将待运算数据的winograd正变换拆解为求和运算之前，将所述待运算数据作为待量化数据进行量化处理；

[0320] 在进行对位乘操作之前，将待运算数据中每个数据的winograd正变换结果作为待量化数据进行量化处理；

[0321] 在进行winograd逆变换拆解为求和运算之前，将所述对位乘结果作为待量化数据进行量化处理。

[0322] 条款A16.根据条款A14所述的装置，所述将待运算数据的winograd正变换拆解为求和运算，并进行计算得到所述待运算数据中每个数据的winograd正变换结果，包括：

[0323] 将所述待运算数据中的每个数据分别拆解为多个第一子张量，对所述待运算数据中的每个数据的多个第一子张量进行winograd正变换并求和得到所述待运算数据中每个数据的winograd正变换结果，

[0324] 其中，每个数据拆分出的多个第一子张量的个数与对应的数据中不为0元素的个数相同，每个第一子张量中有一个元素与对应数据中的对应位置的元素相同、其他元素均为0。

[0325] 条款A17.根据条款A14所述的装置，所述将对所述对位乘结果的winograd逆变换拆解为求和运算，得到所述winograd卷积结果，包括：

[0326] 将所述对位乘结果拆解为多个第二子张量，对所述多个第二子张量进行winograd逆变换并求和，得到所述待运算数据的winograd卷积结果；

[0327] 其中，所述多个第二子张量的个数与所述对位乘结果中不为0元素的个数相同，所述多个第二子张量中的每个第二子张量中有一个元素与所述对位乘结果中的对应位置的元素相同、其他元素均为0。

[0328] 条款A18.根据条款A13所述的装置，所述统计结果包括以下任一种：每种待量化数据中的绝对值最大值、每种待量化数据中的最大值和最小值的距离的二分之一，

[0329] 所述量化参数包括点位置参数、缩放系数和偏移量中的一种或多种，

[0330] 其中，所述绝对值最大值是每种待量化数据中的最大值或最小值的绝对值。

[0331] 条款A19.根据条款A18所述的装置，所述缩放系数是根据所述点位置参数、所述统计结果、所述数据位宽确定的。

[0332] 条款A20.根据条款A18所述的装置，所述偏移量是根据每种待量化数据的统计结果确定的。

[0333] 条款A21.根据条款A18所述的装置，所述点位置参数是根据所述统计结果和所述数据位宽确定的。

[0334] 条款A22.根据条款A18所述的装置，所述装置还包括：

[0335] 位宽调整模块，根据所述数据位宽对应的量化误差，对所述数据位宽进行调整，以利用调整后的数据位宽确定量化参数，

[0336] 其中，所述量化误差是根据对应层中量化后的数据与对应的量化前的数据确定的。

[0337] 条款A23.根据条款A22所述的装置，所述量化前的数据是在目标迭代间隔内的权值更新迭代过程中涉及的待量化数据；

[0338] 其中，所述目标迭代间隔包括至少一次权值更新迭代，且同一目标迭代间隔内量

化过程中采用相同的所述数据位宽。

[0339] 条款A24.根据条款A14所述的装置，所述待运算数据包括输入神经元、权值和梯度中的至少一种。

[0340] 条款A25.一种人工智能芯片，所述芯片包括如条款A13至条款A24中任意一项所述的数据处理装置。

[0341] 条款A26.一种电子设备，所述电子设备包括如条款A25所述的人工智能芯片。

[0342] 条款A27.一种板卡，所述板卡包括：存储器件、接口装置和控制器件以及如条款A25所述的人工智能芯片；

[0343] 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；

[0344] 所述存储器件，用于存储数据；

[0345] 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；

[0346] 所述控制器件，用于对所述人工智能芯片的状态进行监控。

[0347] 条款A28.根据条款A27所述的板卡，

[0348] 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；

[0349] 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；

[0350] 所述接口装置为：标准PCIE接口。

[0351] 条款A29.一种电子设备，包括：

[0352] 处理器；

[0353] 用于存储处理器可执行指令的存储器；

[0354] 其中，所述处理器被配置为调用所述存储器存储的指令，以执行条款A1至条款A12中任意一项所述的方法。

[0355] 条款A30.一种计算机可读存储介质，其上存储有计算机程序指令，所述计算机程序指令被处理器执行时实现条款A1至条款A12中任意一项所述的方法。

[0356] 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。

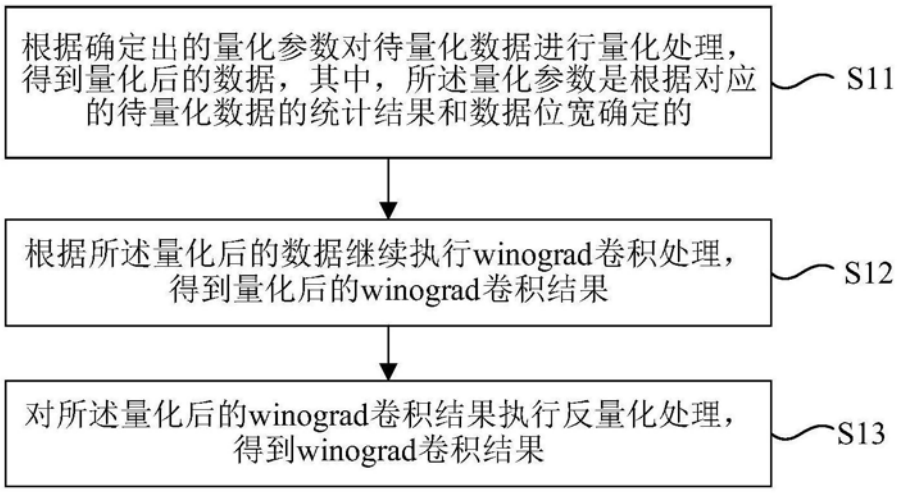

根据确定出的量化参数对待量化数据进行量化处理，得到量化后的数据，其中，所述量化参数是根据对应的待量化数据的统计结果和数据位宽确定的
S11
根据所述量化后的数据继续执行winograd卷积处理，得到量化后的winograd卷积结果
S12
对所述量化后的winograd卷积结果执行反量化处理，得到winograd卷积结果
S13

图1

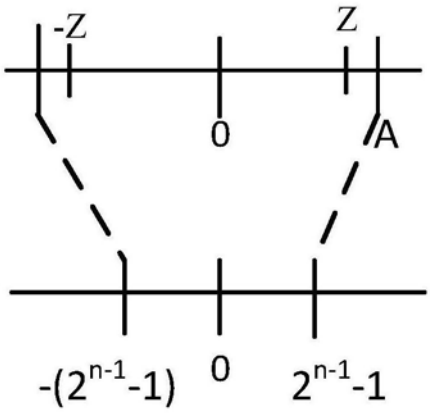

-Z
Z
0
A
-(2^n-1-1)
0
2^n-1-1

图2

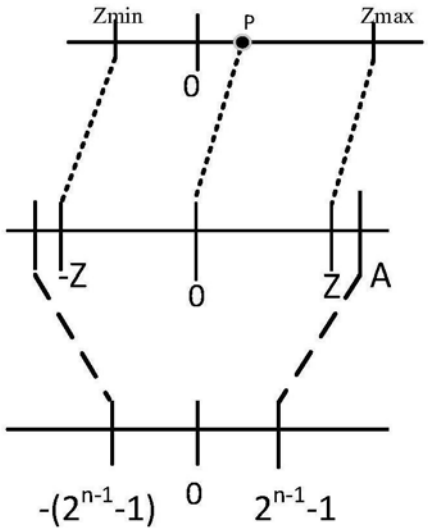

Zmin
P
Zmax
0
-Z
0
Z
A
-(2^n-1-1)
0
2^n-1-1

图3

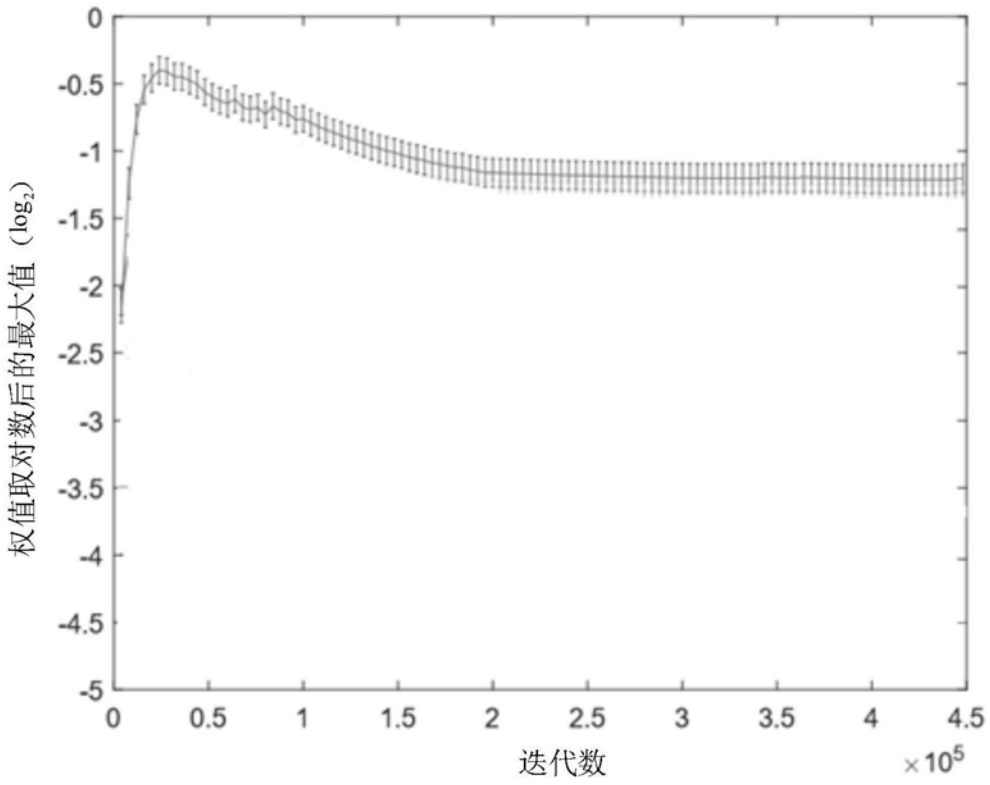

权值取对数后的最大值（log2）
迭代数
×10^5
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
-4
-4.5
-5
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5

图4a

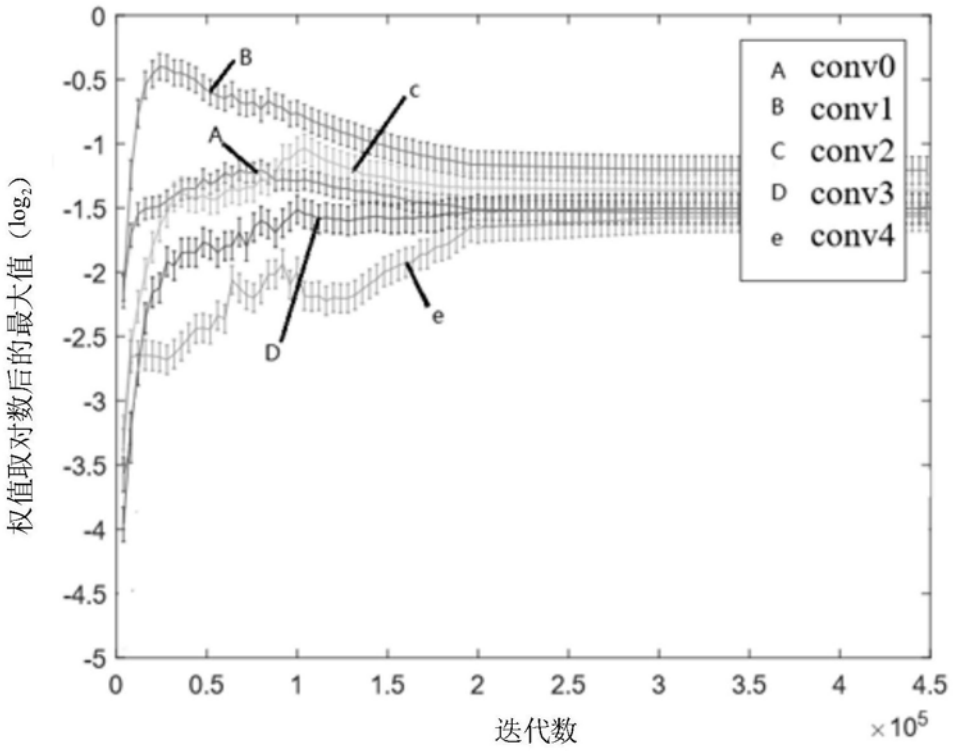

0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
-4
-4.5
-5
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
迭代数
×10^5
权值取对数后的最大值（log2）
A conv0
B conv1
C conv2
D conv3
e conv4
A
B
C
D
e

图4b

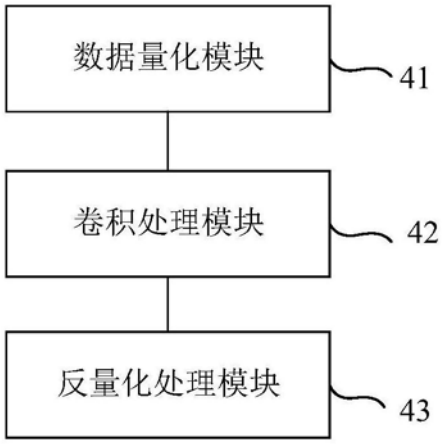

数据量化模块
41
卷积处理模块
42
反量化处理模块
43

图5

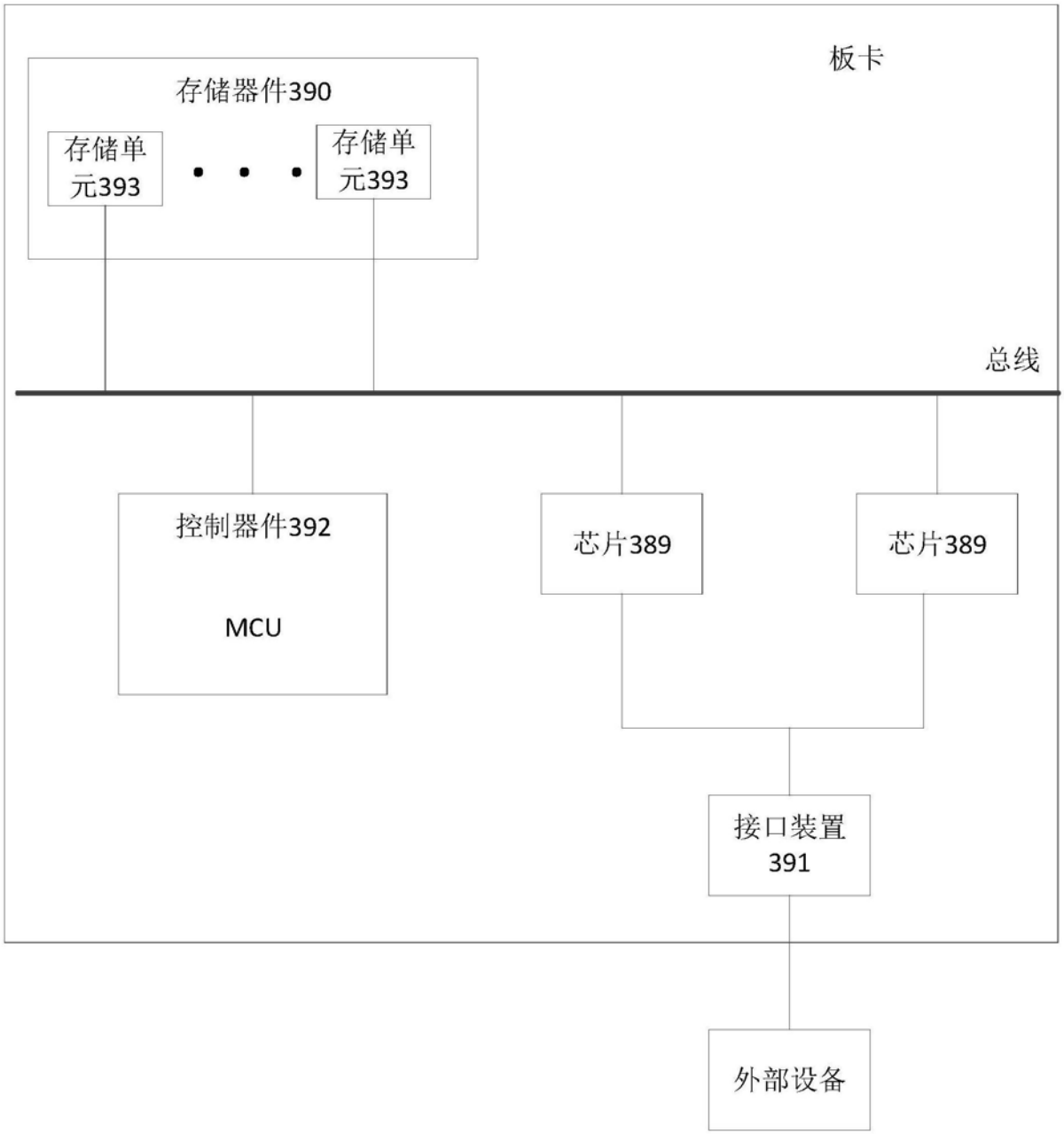

板卡
存储器件390
存储单元393
存储单元393
总线
控制器件392
MCU
芯片389
芯片389
接口装置391
外部设备

图6

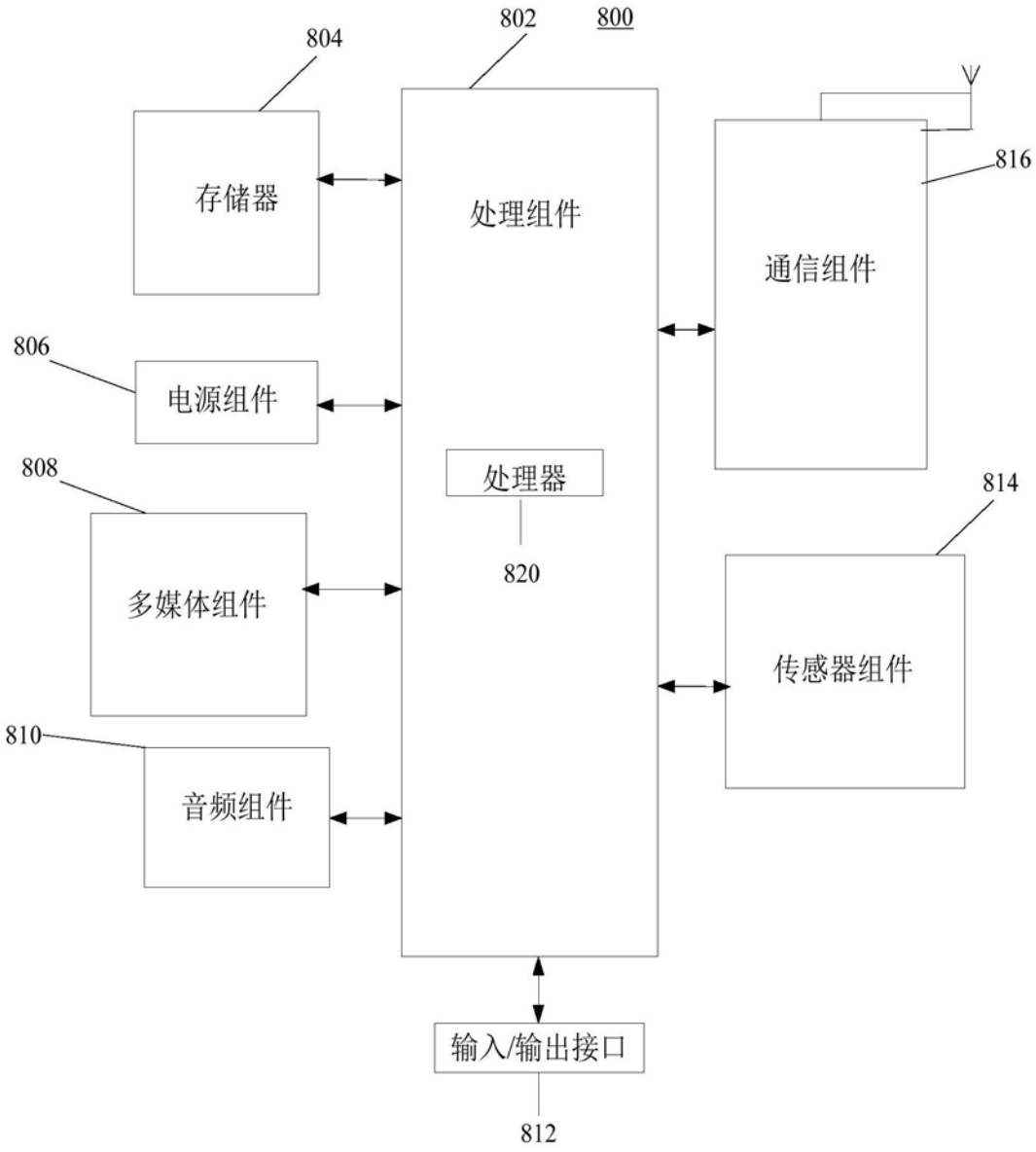

802
800
804
存储器
处理组件
816
通信组件
806
电源组件
处理器
820
808
多媒体组件
814
传感器组件
810
音频组件
输入/输出接口
812

图7

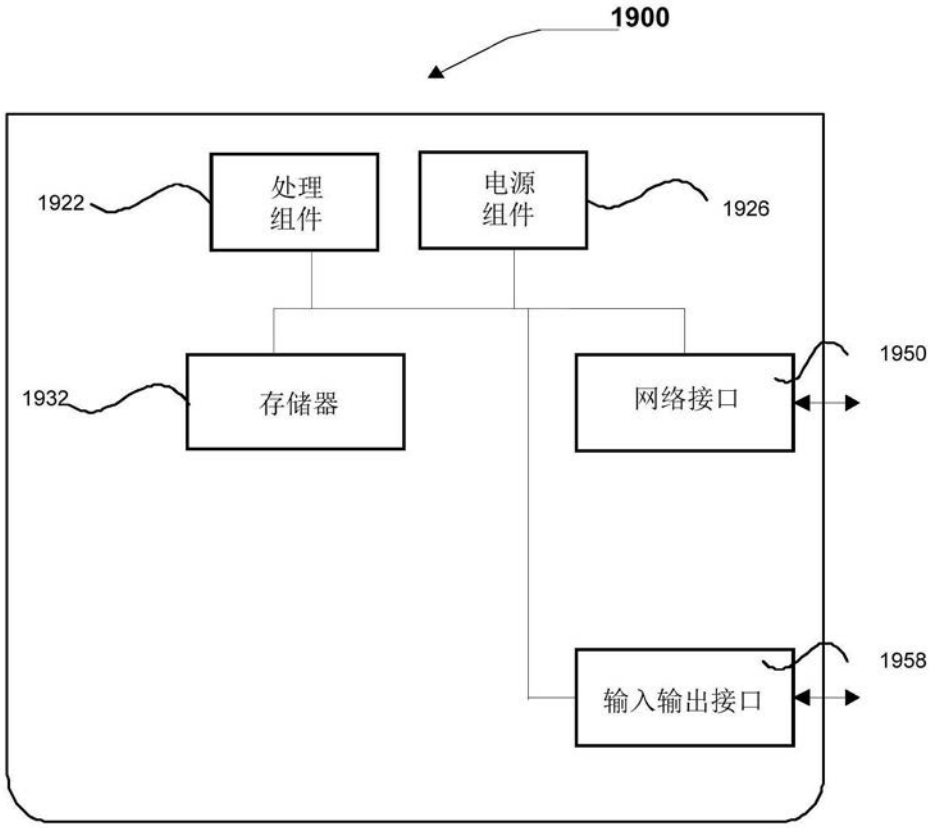


图8