



## (12) 发明专利申请

(10) 申请公布号 CN 119514620 A

(43) 申请公布日 2025. 02. 25

(21) 申请号 202411160917.8

(22) 申请日 2024.08.22

(30) 优先权数据

10-2023-0111513 2023.08.24 KR

(71) 申请人 三星电子株式会社

地址 韩国京畿道

申请人 高丽大学校产学协力团

(72) 发明人 朴相优 徐泰源

(74) 专利代理机构 北京市立方律师事务所

11330

专利代理师 谢玉斌

(51) Int. Cl.

G06N 3/063 (2023.01)

G06N 3/045 (2023.01)

G06N 3/082 (2023.01)

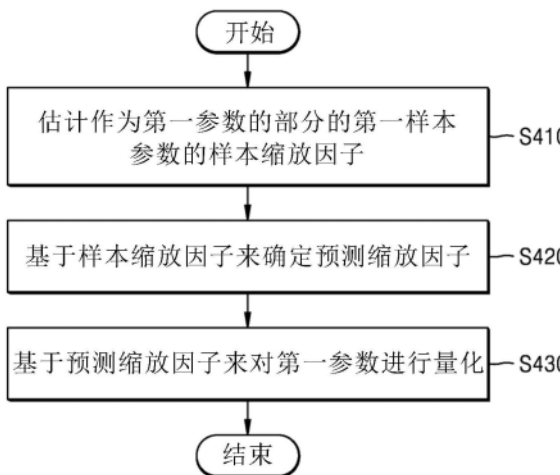
权利要求书2页 说明书12页 附图7页

(54) 发明名称

用于人工神经网络的量化方法、量化系统和  
量化设备

(57) 摘要

提供了用于人工神经网络的量化方法、量化系统和量化设备。所述用于人工神经网络的量化方法可以包括：估计第一样本参数的样本缩放因子，所述第一样本参数是所述人工神经网络内的第一参数中的一部分第一参数；基于所述样本缩放因子来确定预测缩放因子；以及基于所述预测缩放因子来对所述第一参数进行量化。



1. 一种用于人工神经网络的量化方法,所述量化方法包括:  
估计第一样本参数的样本缩放因子,所述第一样本参数是所述人工神经网络内的第一参数中的一部分第一参数;  
基于所述样本缩放因子来确定预测缩放因子;以及  
基于所述预测缩放因子来对所述第一参数进行量化。
2. 根据权利要求1所述的量化方法,其中,所述第一参数包括被输入到所述人工神经网络内的第一层的输入变量和所述第一层的权重。
3. 根据权利要求1所述的量化方法,其中,确定所述预测缩放因子包括:基于所述样本缩放因子的指数移动平均值来确定所述预测缩放因子。
4. 根据权利要求1所述的量化方法,其中,确定所述预测缩放因子包括:基于所述样本缩放因子的算术均值来确定所述预测缩放因子。
5. 根据权利要求1所述的量化方法,其中,确定所述预测缩放因子包括:基于所述样本缩放因子的最大值或最小值来确定所述预测缩放因子。
6. 根据权利要求1所述的量化方法,其中,对所述第一参数进行量化包括:  
通过仿射量化或缩放量化对所述第一参数进行量化。
7. 根据权利要求6所述的量化方法,其中,对所述第一参数进行量化进一步包括:  
通过使用所述预测缩放因子而不是所述第一参数各自的缩放因子来对所述第一参数进行量化。
8. 一种用于人工神经网络的量化系统,所述量化系统包括:  
至少一个处理器;以及  
存储介质,所述存储介质被配置为存储能由所述至少一个处理器运行以执行所述人工神经网络的量化过程的命令,  
其中,所述人工神经网络的所述量化过程包括:  
估计第一样本参数的样本缩放因子,所述第一样本参数是所述人工神经网络内的第一参数中的一部分第一参数;  
基于所述样本缩放因子来确定预测缩放因子;以及  
基于所述预测缩放因子来对所述第一参数进行量化。
9. 根据权利要求8所述的量化系统,其中,所述第一参数包括被输入到所述人工神经网络内的第一层的输入变量和所述第一层的权重。
10. 根据权利要求8所述的量化系统,其中,确定所述预测缩放因子包括:基于所述样本缩放因子的指数移动平均值来确定所述预测缩放因子。
11. 根据权利要求8所述的量化系统,其中,确定所述预测缩放因子包括:基于所述样本缩放因子的算术均值来确定所述预测缩放因子。
12. 根据权利要求8所述的量化系统,其中,确定所述预测缩放因子包括:基于所述样本缩放因子的最大值或最小值来确定所述预测缩放因子。
13. 根据权利要求8所述的量化系统,其中,对所述第一参数进行量化包括:通过仿射量化或缩放量化对所述第一参数进行量化。
14. 根据权利要求8所述的量化系统,其中,所述至少一个处理器被配置为:随着初始数据被输入到所述至少一个处理器而实时地执行所述人工神经网络的所述量化过程。

15. 一种用于人工神经网络的量化设备,所述量化设备包括:

缩放因子估计器,所述缩放因子估计器被配置为估计第一样本参数的样本缩放因子,所述第一样本参数是所述人工神经网络内的第一参数中的一部分第一参数,并且所述缩放因子估计器被配置为基于所述样本缩放因子来确定预测缩放因子;以及

量化器,所述量化器被配置为基于所述预测缩放因子来对所述第一参数进行量化。

16. 根据权利要求15所述的量化设备,其中,所述第一参数包括被输入到所述人工神经网络内的第一层的输入变量和所述第一层的权重。

17. 根据权利要求15所述的量化设备,其中,所述预测缩放因子是基于所述样本缩放因子的指数移动平均值或算术均值而确定的。

18. 根据权利要求15所述的量化设备,其中,所述预测缩放因子是基于所述样本缩放因子的最大值或最小值而确定的。

19. 根据权利要求15所述的量化设备,其中,所述量化器被配置为:通过仿射量化或缩放量化对所述第一参数进行量化。

20. 根据权利要求19所述的量化设备,其中,所述量化器被配置为:通过使用所述预测缩放因子而不是所述第一参数各自的缩放因子来对所述第一参数进行量化。

## 用于人工神经网络的量化方法、量化系统和量化设备

### 技术领域

[0001] 本发明构思涉及人工神经网络,更具体地,涉及用于人工神经网络的量化方法、量化设备和量化系统。

### 背景技术

[0002] 人工神经网络可以指计算装置或由计算装置执行以实现一组互连的人工神经元(或神经元模型)的方法。人工神经元可以通过对输入数据执行简单运算来生成输出数据,并且该输出数据可以被发送到其他人工神经元。作为人工神经网络的示例,深度神经网络或深度学习可以具有多层结构。

[0003] 当在深度学习推理操作中使用动态量化时,模型的每个层的输入可能需要缩放因子(scale factor)。当通过为每个层的每个输入获得缩放因子来执行量化操作时,深度学习推理可能需要大量计算。

### 发明内容

[0004] 本发明构思提供了用于人工神经网络的量化方法、量化系统和量化设备,在其中可以实现人工神经网络的高准确度和低计算复杂度。

[0005] 根据一个或多个实施例,提供了一种用于人工神经网络的量化方法,所述量化方法包括:估计第一样本参数的样本缩放因子,所述第一样本参数是所述人工神经网络内的第一参数中的一部分第一参数;基于所述样本缩放因子来确定预测缩放因子;以及基于所述预测缩放因子来对所述第一参数进行量化。

[0006] 根据一个或多个实施例,提供了一种用于人工神经网络的量化系统,所述量化系统包括:至少一个处理器;以及存储介质,所述存储介质被配置为存储能由所述至少一个处理器运行以执行所述人工神经网络的量化过程的命令,其中,所述人工神经网络的所述量化过程包括:估计第一样本参数的样本缩放因子,所述第一样本参数是所述人工神经网络内的第一参数中的一部分第一参数;基于所述样本缩放因子来确定预测缩放因子;以及基于所述预测缩放因子来对所述第一参数进行量化。

[0007] 根据一个或多个实施例,提供了一种用于人工神经网络的量化设备,所述量化设备包括:缩放因子估计器,所述缩放因子估计器被配置为估计第一样本参数的样本缩放因子,所述第一样本参数是所述人工神经网络内的第一参数中的一部分第一参数,并且所述缩放因子估计器被配置为基于所述样本缩放因子来确定预测缩放因子;以及量化器,所述量化器被配置为基于所述预测缩放因子来对所述第一参数进行量化。

### 附图说明

[0008] 根据以下结合附图进行的详细描述,将更清楚地理解示例实施例,在附图中:

[0009] 图1是图示了根据示例实施例的量化系统的框图;

[0010] 图2是图示了根据示例实施例的人工神经网络的示例的视图;

- [0011] 图3A和图3B是图示了根据示例实施例的一种量化方法的视图；
- [0012] 图4是图示了根据示例实施例的用于人工神经网络的量化方法的流程图；
- [0013] 图5是图示了根据示例实施例的用于人工神经网络的量化设备的框图；
- [0014] 图6是图示了根据示例实施例的图5的量化设备内部的处理装置的框图；
- [0015] 图7是图示了根据示例实施例的图5的量化设备内部的量化与缩放因子估计(Quantization with Scale factor Estimation, QSE)单元的框图；
- [0016] 图8是图示了根据示例实施例的计算系统的框图；以及
- [0017] 图9是图示了根据示例实施例的便携式计算系统的框图。

## 具体实施方式

[0018] 在下文中,将参考附图详细地描述本发明构思的示例实施例。

[0019] 图1是图示了根据示例实施例的量化系统100的框图。人工神经网络可以指试图模拟动物大脑中的生物神经网络的计算系统。在人工神经网络中,可以通过考虑多个样本(或示例)来学习任务的执行,这不同于诸如根据预定义条件来执行任务的基于规则的编程的经典算法。人工神经网络可以具有其中人工神经元(或神经元)彼此连接的结构,并且各神经元之间的连接可以被称为神经键(synapse)。神经元可以处理接收到的信号,并且可以通过神经键将处理的信号发送到另一神经元。神经元的输出可以被称为激活。神经元和/或神经键可以具有可变权重,并且由神经元处理的信号的影响可以根据权重而增加或减小。另外,应用了与单独神经元相关联的偏置。

[0020] 深度神经网络或深度学习架构可以具有层结构,并且特定层的输出可以是后续层的输入。在这样的多层结构中,可以根据多个样本对每个层进行训练。人工神经网络(诸如深度神经网络)可以由分别对应于人工神经元的多个处理节点来实现,并且可能需要高计算复杂度来获得令人满意的结果,例如高准确度结果。因此,可能需要许多计算资源。

[0021] 为了降低计算复杂度,可以对人工神经网络进行量化。量化可以指将输入值映射到数目比输入值的数目小的值的过程,诸如通过四舍五入将实数映射为整数。在人工神经网络中,量化可以包括将浮点十进制神经网络转换成整数神经网络的过程。例如,在人工神经网络中,量化可以用在激活、层的权重等中。浮点十进制数可以包括符号、指数和有效数,其中整数可以包括整数部分。在一些实施例中,整数的整数部分可以包括符号位。参考图1,输入数据集IN用浮点十进制数表达,并且经历量化过程的输出数据集OUT可以用整数表达。使用浮点十进制数的人工神经网络可以具有高准确度和高计算复杂度,然而使用整数的人工神经网络可以具有降低的准确度和更低的计算复杂度。

[0022] 在人工神经网络中,由于结果的准确度与计算复杂度之间的权衡关系,人工神经网络的量化可以导致准确度减小,并且准确度降低的程度可以取决于量化的方法。在下文中,如下面参考附图描述的,根据示例实施例的量化系统100可以在使准确度的降低最小化时根据要求提供量化,因此,可以提供复杂度降低同时性能足够高的量化神经网络。

[0023] 量化系统100可以是根据示例实施例执行量化的任何类型的系统,并且也可以被称为量化设备。例如,量化系统100可以是包括至少一个处理器和至少一个存储器的计算系统。作为非限制性示例,量化系统100可以是诸如膝上型计算机、智能电话等的移动计算系统以及诸如台式计算机和服务器等的固定计算系统。如图1所示,量化系统100可以包括估

计器120和量化器140,并且估计器120和量化器140中的每一者可以由通过逻辑合成实现的逻辑块、由处理器执行的软件块或它们的组合来实现。在一些实施例中,估计器120和量化器140中的每一者可以是作为由处理器运行的多个命令的集合的过程,或者可以被存储在处理器可访问的存储器中。

[0024] 在本文中,缩放因子估计器120和量化器140可以各自是包括以下一者或更多者的模拟电路和/或数字电路:逻辑门、集成电路、微处理器、微控制器、存储器电路、无源电子部件、有源电子部件、光部件等,并且可以被配置为运行软件和/或固件以执行上述对应的功能或操作。

[0025] 参考图1,量化系统100可以接收输入数据集IN,并且可以生成量化的输出数据集OUT。输入数据集IN可以包括多个输入数据,并且量化的输出数据集OUT可以包括多条输出数据。如图1所示,作为输入数据集IN的多条矩阵型数据可以被输入到根据示例实施例的量化系统100,并且量化系统100可以输出多条矩阵型数据作为量化的输出数据集OUT。根据示例实施例的多条数据可以包括从图像中提取的每个像素的多条数据。构成输入数据集IN的每条输入数据可以包括浮点十进制数。构成要通过量化系统100输出的输出数据集OUT的每条输出数据可以包括整数。如下所述,在根据示例实施例的量化系统100中,可以省略用于在量化操作中为所有输入确定缩放因子的计算,因此,总计算量可以减少。

[0026] 估计器120可以接收输入数据集IN,并且可以确定诸如激活、权重等的预测缩放因子,以将该预测缩放因子提供给量化器140。为了对浮点十进制数型变量进行量化,可能需要缩放因子。缩放因子可以指示用于将正在被量化的数据的值的范围映射到量化范围的值,该量化范围对应于能够由量化过程中使用的位数表示的最大值和最小值。下面将参考图3A和图3B详细地提供缩放因子的详细描述。对变量进行量化的过程可以被分类成动态量化和静态量化。根据示例实施例的量化系统100可以通过使用动态量化方法来对多条数据进行量化,在所述动态量化方法中,人工神经网络中的权重、激活等在学习之后的推理步骤中被实时地量化。在本说明书中为了方便,人工神经网络中的诸如权重、激活等的待量化的多条数据可以被称为参数。当针对待量化的所有参数计算或估计缩放因子时,缩放因子可以具有高准确度,但是可能需要过量计算。如下所述,根据示例实施例的估计器120可以针对待量化的参数的部分估计缩放因子,并且可以从所估计的缩放因子确定预测缩放因子,以将所确定的预测缩放因子提供给量化器140。也就是说,估计器120可以仅针对待量化的参数的部分估计缩放因子,并且跳过针对第一参数的其他部分估计缩放因子。由于量化系统100可以仅对参数的部分而非所有参数估计缩放因子,基于从参数的部分估计的缩放因子来确定预测缩放因子,并且使用该预测缩放因子来对于参数执行量化,所以可以省略用于在深度学习推理期间的量化操作中为所有输入确定缩放因子的计算,因此,总计算量可以减少。

[0027] 量化器140可以从估计器120接收与参数相对应的预测缩放因子,并且可以基于预测缩放因子来对参数进行量化,以生成量化的输出数据集OUT。下面将参考图7等详细地描述量化器140的详细量化过程。

[0028] 图2是图示了根据示例实施例的人工神经网络的示例的视图。

[0029] 参考图2,人工神经网络可以具有包括输入层L1、隐藏层(L2、L3)和输出层Ln(其中n是3或更大的自然数)的结构,并且可以基于所接收到的输入数据(例如, $I_1$ 、 $I_2$ )来执行运

算,以及可以基于运算的结果来生成输出数据(例如, $0_1$ 、 $0_2$ )。

[0030] 人工神经网络可以是包括一个或更多个隐藏层的深度神经网络或n层神经网络。例如,如图2所示,人工神经网络可以是包括输入层L1、隐藏层L2和L3以及输出层Ln的深度神经网络。深度神经网络可以包括卷积神经网络(CNN)、循环神经网络(RNN)、深度信念网络、受限玻尔兹曼机等,并且示例实施例不限于此。

[0031] 当人工神经网络具有深度神经网络(DNN)结构时,人工神经网络包括可以借以提取有效信息的更多层,使得人工神经网络可以处理比在根据现有技术的其他类型的人工神经网络中更复杂的数据集。人工神经网络也可以包括具有与图2所示的结构不同的各种结构的层。

[0032] 包括在人工神经网络中的层L1至Ln中的每一者可以包括多个人工节点,该多个人工节点也被称为神经元、单元或类似术语。例如,如图2所示,输入层L1可以包括两个节点ND,并且隐藏层L2和L3可以包括三个节点ND。然而,这仅是示例,并且包括在人工神经网络中的每个层可以包括各种数目的节点ND,并且每个层中包括的节点ND的数目可以彼此不同。

[0033] 包括在人工神经网络中的每个层中包括的节点可以彼此连接以交换数据。例如,一个节点ND可以从其他节点ND接收数据以执行计算,并且可以将计算的结果输出到其他节点ND。

[0034] 每个节点ND的输入和输出可以被称为激活。激活可以是一个节点ND的输出值,并且可以是下一层中包括的节点ND的输入值。每个节点ND可以基于从前一层中包括的节点ND接收到的激活和权重来确定它自己的激活。权重是用于计算每个节点ND中的激活的网络参数,并且可以是分配给节点ND之间的连接关系的值。例如,在第二层L2中,节点ND可以基于从前一层L1接收到的激活( $a_{11}$ 、 $a_{12}$ )、权重( $w_{21}$ 、 $w_{22}$ 、 $w_{23}$ 、 $w_{24}$ 、 $w_{25}$ 、 $w_{26}$ )和偏置( $b_{21}$ 、 $b_{22}$ 、 $b_{23}$ )来确定它们自己的激活。每个节点ND可以是接收输入并且输出激活的计算单元,并且可以执行输入-输出映射。

[0035] 人工神经网络可以包括各层之间的激活函数。激活函数可以将前一层的输出转换成下一层的输入。例如,激活函数可以是非线性函数,诸如整流线性单元(ReLU)、参数化整流线性单元(PReLU)、双曲正切(tanh)、sigmoid函数,并且可以在第二层L2与第三层L3之间非线性地转换第二层L2的输出。激活可以通过将激活函数应用于从前一层接收到的激活的加权和而获得的值。

[0036] 随后,参考图1和图2,待由量化系统100量化的对象可以包括人工神经网络的激活和权重。例如,从第一层L1接收到的被输入到第二层L2的激活( $a_{11}$ 、 $a_{12}$ )以及权重( $w_{21}$ 、 $w_{22}$ 、 $w_{23}$ 、 $w_{24}$ 、 $w_{25}$ 、 $w_{26}$ )可以是浮点十进制数格式的数据。量化系统100可以通过上述量化过程将激活( $a_{11}$ 、 $a_{12}$ )和权重( $w_{21}$ 、 $w_{22}$ 、 $w_{23}$ 、 $w_{24}$ 、 $w_{25}$ 、 $w_{26}$ )量化成整数格式的数据。加权和可以使用量化的激活和权重来计算,因此深度学习模型的大小可以减小并且计算量可以减少。随后,随着激活函数被应用于所计算出的加权和的值,从第二层L2输入到第三层L3的激活( $a_{21}$ 、 $a_{22}$ 、 $a_{23}$ )可以再次是浮点十进制数格式的数据。也就是说,待由量化系统100量化的对象可以是特定层的参数或者可以是与每个层相对应的参数。

[0037] 图3A和图3B是图示了根据示例实施例的一种量化方法的视图。

[0038] 具体地,图3A图示了仿射量化(Affine quantization)方法,并且图3B图示了缩放

量化(scale quantization)方法。

[0039] 参考图3A,在仿射量化方法中,数据在被量化之前的最大值和最小值可以分别线性地对应于待量化的数据格式的最大值或最小值。例如,如图3A所示,数据在被量化之前的最大值( $\alpha$ )可以是7,并且数据在被量化之前的最小值( $\beta$ )可以是-5。当数据被量化成8个位的整数格式时,量化的数据可以具有范围为-128至127的值。根据仿射量化方法,数据在被量化之前的最大值( $\alpha$ )可以对应于127,并且数据在被量化之前的最小值( $\beta$ )可以对应于-128。最大值( $\alpha$ )与最小值( $\beta$ )之间的数据值可以线性地对应于介于127与-128之间的数据。因此,当数据在被量化之前的最大值( $\alpha$ )和数据在被量化之前的最小值( $\beta$ )具有不同值时,被量化之前的值0可以对应于在被量化之后不是0的另一值。在线性地对应值的过程中,数据的值在被量化之后的范围相对于数据的值在被量化之前的范围的比率可以是缩放因子。也就是说,在仿射量化方法中,缩放因子可以被计算为以下[式1]。

[0040] [式1]

$$[0041] \quad SF(\text{缩放因子}) = \frac{2^b - 1}{\alpha - \beta}$$

[0042] 在[式1]中,b表示整数数据在量化之后的位数。参考图3A所图示的值,图3A的仿射量化方法的缩放因子可以是 $\frac{255}{12}$ 。

[0043] 参考图3B,在缩放量化方法中,类似于仿射量化方法,数据在被量化之前的最大值和最小值可以分别线性地对应于待量化的数据格式的最大值和最小值。然而,在缩放量化方法中,为了使0在被量化之前的值对应于0在被量化之后的值,可以满足关系( $\alpha = -\beta$ )。也就是说,在缩放量化方法中,可以获得数据在被量化之前的绝对值的最大值( $\alpha$ ),并且 $\alpha$ 和 $-\alpha$ 可以分别线性地对应于待量化的数据格式的最大值和最小值。在线性地对应值的过程中,数据的值在被量化之后的范围相对于数据的值在被量化之前的范围的比率可以是缩放因子。也就是说,在缩放量化方法中,缩放因子可以被计算为以下[式2]。

[0044] [式2]

$$[0045] \quad SF = \frac{2^{b-1} - 1}{\alpha}$$

[0046] 在[式2]中,b表示整数数据在被量化之后的位数。参考图3B所图示的值,图3B的缩放量化方法的缩放因子可以是 $\frac{127}{7}$ 。

[0047] 在图3A和图3B的两种量化方法中,缩放因子可以乘以被量化之前的数据,并且通过乘法获得的值可以位于数据在被量化之后的范围内。量化方法可以包括以整数格式表达将缩放因子乘以待量化的数据的结果的舍入运算以及校正超出范围的值的限幅运算,如下面参考图7详细地描述的。在图3A和图3B的两种量化方法中,可能需要计算数据的最大值和最小值的过程来计算或估计缩放因子。当输入各种数据集并且针对所有数据集计算缩放因子时,可能需要过量计算。如下所述,根据示例实施例,可以省略用于在深度学习推理期间的量化操作中为所有输入确定缩放因子的计算,因此总计算量可以减少。

[0048] 图4是图示了根据示例实施例的用于人工神经网络的量化方法的流程图。

[0049] 例如,图4的量化方法可以通过使用图1的量化系统100来执行。在下文中,将参考

图1对图4进行描述。

[0050] 参考图4,在操作S410中,可以执行估计第一样本参数的样本缩放因子的操作。第一样本参数可以是第一参数的部分。第一参数可以包括被输入到第一层的输入变量或第一层的权重,这些输入变量或权重将被量化。输入到第一层的输入变量可以是来自第一层的前一层的激活输出。

[0051] 作为第一参数的部分的第一样本参数可以是用于确定将在下面描述的预测缩放因子的参数。第一样本参数可以通过对输入的第一参数的部分(不是预定义样本)进行采样而选择的。例如,当输入的第一参数包括A个数据集时,第一样本参数可以包括输入的第一参数的前B个数据集(其中B是小于自然数A的自然数)。也就是说,在示例实施例中,第一样本参数可以指第一参数当中输入的一系列(例如,连续的)第一参数。输入的第一参数的前B个数据集可以是第一样本参数,并且可以对输入的第一参数的前B个数据集执行估计或计算关于每个第一样本参数的缩放因子的操作。与第一个样本参数相对应的缩放因子可以通过使用上面参考图1至图3描述的方法来估计的。例如,当输入数据集IN如图1所示作为第一样本参数被输入时,可以获得数据集中的最大值或最小值,并且可以据此计算或估计缩放因子,如上面参考图3描述的那样。为了获得数据集中的最大值或最小值,可以利用比较器或者可以执行比较操作,如下面参考图7详细地描述的那样。

[0052] 参考图4,在操作S420中,可以执行基于样本缩放因子来确定预测缩放因子的操作。在上一操作S410中,在针对第一样本参数估计样本缩放因子之后,可以基于样本缩放因子来确定可以表示所有第一参数的预测缩放因子。如上所述,当针对所有参数计算缩放因子时,可能需要过量计算。因此,在本公开的量化方法中,可以在量化计算中利用预测缩放因子。在基于样本缩放因子来确定预测缩放因子的方法中,可以利用各种式,并且可以根据数据集的特性或关系来选择不同式。根据示例实施例的确定预测缩放因子的方法可以使用以下[式3]。

[0053] [式3]

$$[0054] \quad PSF(\text{预测缩放因子}) = \frac{PSF + SF}{2}$$

[0055] [式3]是将由计算机和/或程序执行的计算式,并且是指示当样本缩放因子SF被顺序地输入到量化器140时预测缩放因子PSF的计算方法的计算式。在[式3]中,具体地,可以通过反映指数移动平均值来计算预测缩放因子。[式3]示出了在预测缩放因子PSF的计算方法中比例系数为1/2的情况,并且实施例不限于[式3]。基于[式3]的另一示例公式被示出在下面的式4中。

[0056] [式4]

$$[0057] \quad PSF = (1 - k) \cdot PSF + k \cdot SF$$

[0058] [式4]是将由计算机和/或程序执行的计算式,并且是指示当样本缩放因子SF被顺序地输入到量化器140时预测缩放因子PSF的计算方法的计算式。k是介于0与1之间的实数,并且k值越大,前一观察值的影响可以降低得越大。当k是1/2时,[式3]成立。

[0059] 根据另一示例实施例的确定预测缩放因子的方法可以使用以下[式5]。

[0060] [式5]

$$[0061] \quad PSF = \frac{SF}{(\text{样本数})}$$

[0062] [式5]是将由计算机和/或程序执行的计算式,并且是指示当样本缩放因子SF被顺序地输入到量化器140时预测缩放因子PSF的计算方法的计算式。在[式5]中,具体地,可以通过反映算术均值来计算预测缩放因子。

[0063] 根据另一示例实施例的确定预测缩放因子的方法可以使用以下[式6]。

[0064] [式6]

$$[0065] \quad PSF = \min(PSF, SF)$$

[0066] [式6]是将由计算机和/或程序执行的计算式,并且是指示当样本缩放因子SF被顺序地输入到量化器140时预测缩放因子PSF的计算方法的计算式。在[式6]中,具体地,可以通过反映样本缩放因子的最小值来计算预测缩放因子。

[0067] 根据另一示例实施例的确定预测缩放因子的方法可以使用以下[式7]。

[0068] [式7]

$$[0069] \quad PSF = \max(PSF, SF)$$

[0070] [式7]是将由计算机和/或程序执行的计算式,并且是指示当样本缩放因子SF被顺序地输入到量化器140时预测缩放因子PSF的计算方法的计算式。在[式7]中,具体地,可以通过反映样本缩放因子的最大值来计算预测缩放因子。

[0071] 诸如[式3]至[式7]的式可以通过针对按时间顺序输入的样本参数顺序地计算样本缩放因子、并且将顺序地计算出的样本缩放因子累加地反映在预测缩放因子值中来计算的。在这种情况下,首先计算(或估计)的样本缩放因子可以被确定为初始预测缩放因子。另外,在计算出与样本参数相对应的样本缩放因子之后,也可以基于所有样本缩放因子来计算预测缩放因子值。

[0072] 通过诸如[式3]至[式7]的式,在操作S420中,可以执行基于样本缩放因子来确定预测缩放因子的操作。然而,这仅是示例式,并且基于样本缩放因子来确定预测缩放因子的方法不限于此。例如,从样本缩放因子计算出的预测缩放因子值可以通过向诸如[式3]至[式7]的式添加或乘以适当的系数被调整。另外,也可以通过组合包括一个或更多个上述式的多个式来计算预测缩放因子的值。通过利用适当的式,可以执行其中可以保证准确度同时减少人工神经网络的计算量的量化操作。

[0073] 参考图4,在操作S430中,可以执行基于预测缩放因子来对第一参数进行量化的操作。量化操作可以由图1的量化器140的操作来执行,并且也可以由将在下面描述的包括图7的配置的量化设备来执行。对第一参数进行量化的方法可以包括上面参考图3描述的仿射量化方法或缩放量化方法。具体地,在上一操作S420中计算出的预测缩放因子可以是所计算出的因子,而不是针对所有参数计算缩放因子的过程,并且可以用在第一参数的量化过程中。不是计算每个第一参数的缩放因子,而是可以使用所计算出的预测缩放因子来对第一参数进行量化。根据上述实施例,通过使用预测缩放因子,可以省略用于在深度学习推理期间的量化操作中为所有输入确定缩放因子的计算,因此,总计算量可以减少。另外,用于基于量化的具有高准确度的人工神经网络来实现人工神经网络的计算资源可以减少,并且人工神经网络的应用范围可以扩展。

[0074] 图5是图示了根据示例实施例的用于人工神经网络的量化设备的框图。

[0075] 参考图5,用于人工神经网络的量化设备500可以包括存储器510和处理器520。在下文中,将参考图1和图4对图5进行描述,并且省略与参考先前附图的描述的冗余描述。

[0076] 量化设备500的存储器510可以存储用于根据示例实施例的人工神经网络的量化的程序,并且可以存储通过量化方法量化的激活或数据。另外,存储器510可以存储图1的量化的输出数据集OUT,并且可以存储在量化过程中生成的偏置或数据。存储器510可以存储与量化的执行相关的一系列命令。存储器510可以与处理器520进行通信以彼此交换数据。

[0077] 处理器520可以是中央处理单元(CPU)、图形处理单元(GPU)和神经处理单元(NPU)中的一者或更多者。然而,这是示例,并且处理器520不限于以上描述。图5所图示的处理器520可以包括数据取出器522、矢量处理单元524和内积阵列526。数据取出器522可以从存储器510接收诸如权重、激活、偏置等的数据,并且可以将该数据发送到矢量处理单元524。矢量处理单元524可以在其中包括量化与缩放因子估计(QSE)单元524\_1。图5的图示是示例,并且QSE单元524\_1可以在矢量处理单元524外部独立地工作。QSE单元524\_1可以获得与诸如输入权重、激活等的参数相对应的预测缩放因子,并且可以基于预测缩放因子来执行量化。下面将参考图7描述QSE单元524\_1的详细操作。

[0078] 矢量处理单元524可以将量化的参数发送到数据取出器522,并且数据取出器522可以将量化的参数发送到内积阵列526。内积阵列526可以基于诸如输入量化的权重、量化的激活、量化的偏置等的数据来计算加权和。图5图示了量化设备100包括N个处理单元526\_1、526\_2、...、526\_N。然而,这是示例,并且量化设备100也可以包括不同数目的处理单元PE。处理单元PE可以计算量化的权重与量化的激活之间的加权和,并且下面将参考图6详细地描述处理单元PE的详细操作。当基于未量化的参数来计算加权和时,内积阵列526需要大量计算,并且在深度学习的推理操作中可能需要更多时间和许多资源。根据示例实施例,可以省略用于在深度学习推理期间的量化操作中为所有输入确定缩放因子的计算,因此,总计算量可以减少。

[0079] 通过内积阵列526输出的诸如激活的数据可以被发送到矢量处理单元524,并且矢量处理单元524可以从数据取出器522接收附加数据以执行下一层的量化操作。另外,矢量处理单元524可以将从内积阵列526发送的输出激活数据发送到存储器510,以便存储输出激活数据。

[0080] 尽管在图5中未示出,但是可以另外地包括用于在其中存储数据的缓冲器或寄存器的配置,并且可以另外地包括诸如加法器、选择或添加数据的多路选择器(multiplexer)的配置。

[0081] 图6是图示了根据示例实施例的图5的量化设备500内部的处理装置的框图。

[0082] 参考图5和图6,包括在设备500中的内积阵列526可以包括多个处理装置PE 526\_1、526\_2、...、526\_N。每个处理装置PE(例如,如图6所图示的526\_N)可以包括多个乘法器MPR和加法器树。处理装置PE 526\_N可以执行以下[式8]所示的计算。

[0083] [式8]

$$[0084] \quad y = \sum_{k=1}^X I_k W_k + Bias$$

[0085] 在[式8]中,  $I$ 表示输入量化的激活,  $W$ 表示输入量化的权重,  $Bias$ 表示偏置, 并且 $y$ 表示输出。处理装置PE 526\_N可以接收 $X$ 个量化的激活、与量化的激活相对应的量化的权重以及偏置。 $X$ 可以表示对其执行计算的层的通道数, 并且对于每层可以具有不同值。另外, 图6图示了存在 $X$ 个乘法器MPR。然而, 这是示例, 并且处理装置PE 526\_N可以根据硬件规格包括不同数目的乘法器MPR。通过乘法器相乘的量化的激活和量化的权重可以连同偏置一起由加法器树相加。结果, 处理装置PE 526\_N可以通过接收 $X$ 个量化的激活、量化的权重和偏置值来计算加权和。

[0086] 尽管在图6中未示出, 但是根据示例实施例的处理装置PE 526\_N可以包括根据实施例的用于存储数据的寄存器或存储器, 并且可以另外地包括用于应用激活函数的部件。

[0087] 图7是图示了根据示例实施例的图5的量化设备内部的QSE单元的框图。

[0088] 参考图5和图7, 量化设备500内部的矢量处理单元524可以包括QSE单元524\_1。QSE单元524\_1可以包括缩放因子估计器710和量化器720。缩放因子估计器710和量化器720可以是图1的估计器120和量化器140的示例, 并且可以执行部分公共功能。关于量化方法的操作, 省略了与图1和图4的描述的冗余描述。

[0089] QSE单元524\_1可以执行诸如激活、权重等的参数的缩放因子估计和量化。QSE单元524\_1可以接收第一参数并且可以输出量化的参数。在这种情况下, 每个参数可以指数据集, 并且每个数据集可以包括 $N$ 个数据。例如, 在具有 $3 \times 3$ 个像素的图像的情况下, 一个图像数据集可以包括与9个像素相对应的9条数据, 并且在第一层中的量化操作中,  $N$ 可以是9。在这种情况下, 每个第一参数可以包括9条数据, 并且QSE单元524\_1可以接收构成一个输入参数的输入数据 $0\_1$ 、 $0\_2$ 、...、和 $0\_N$ , 并且可以输出量化的输出数据 $Q\_1$ 、 $Q\_2$ 、...、和 $Q\_N$ 。量化的输出数据 $Q\_1$ 、 $Q\_2$ 、...、和 $Q\_N$ 可以构成一个量化的输出参数。也就是说, 图7所图示的输入和输出操作是与一个参数(即, 一个数据集)相对应的操作。例如, 当处理1000个图像(每个图像具有 $3 \times 3$ 个像素)时, 相对于一个图像, 第一层可以接收9条输入数据 $0\_1$ 、 $0\_2$ 、...、和 $0\_N$ , 并且可以输出量化的输出数据 $Q\_1$ 、 $Q\_2$ 、...、和 $Q\_N$ 。另外, 可以对1000个图像执行对应数据的输入和输出。

[0090] QSE单元524\_1内部的缩放因子估计器710可以估计第一样本参数的样本缩放因子。第一样本参数可以是第一参数的部分。在实施例中, QSE单元524\_1可以仅针对待量化的第一参数的部分估计缩放因子, 并且跳过针对第一参数的其它部分估计缩放因子。如上面参考图4描述的, 根据示例实施例的第一样本参数可以指第一参数当中的特定数目的第一参数。第一样本参数可以是连续的第一参数, 并且可以是首先出现的一组第一参数。对于每个第一样本参数, 缩放因子估计器710可以通过使用多个比较器CPR来计算样本参数的最大值和最小值, 以针对所对应的样本参数估计样本缩放因子。如上面参考图3A和图3B描述的, 缩放因子估计器710可以基于通过多个比较器CPR计算出的最大值和最小值来计算样本缩放因子。

[0091] 缩放因子估计器710可以包括被配置为在其中存储样本缩放因子和预测缩放因子的寄存器711。缩放因子估计器710可以通过使用上述多个比较器CPR来计算每个第一样本参数的样本缩放因子, 直到 $B$ 个第一样本参数(例如,  $B$ 个数据集)被输入到缩放因子估计器710。可以预设 $B$ 的值。所计算出的样本缩放因子可以被递送到量化器720并且可以被存储在寄存器711中。缩放因子估计器710可以基于诸如[式3]至[式6]的式和所计算出的样本缩放

因子来计算预测缩放因子,如上面参考图4描述的那样。

[0092] 在所有第一样本参数(例如,B个数据集)都被输入到QSE单元524\_1之后,缩放因子估计器710的寄存器711可以计算和存储预测缩放因子。随后,当附加第一参数被输入到QSE单元524\_1时,缩放因子估计器710可以不执行用于通过每个附加第一参数的多条数据之间的比较计算来计算最大值或最小值的计算。也就是说,在预测缩放因子的计算完成之后,缩放因子估计器710可以将存储在寄存器711中的预测缩放因子输出到量化器720,而不根据第一参数的输入来估计缩放因子。可以不对所有参数(例如,所有数据集)执行缩放因子计算,并且可以使用预测缩放因子来执行量化计算。因此,根据示例实施例,可以省略用于在深度学习推理期间的量化操作中为所有输入确定缩放因子的计算,因此,总计算量可以减少并且推理可以被加速。

[0093] 在输入数据0\_1、0\_2、...、和0\_N以及来自缩放因子估计器710的缩放因子SF被输入量化器720之后,量化器720可以输出量化的输出数据Q\_1、Q\_2、...、和Q\_N。在第一样本参数被输入到量化器720的部分中,所计算出的样本缩放因子可以被从缩放因子估计器710输入到量化器720。图7示出了缩放因子SF被输入到量化器720。然而,这是综合表示法,并且在预测缩放因子的计算完成之后,如上所述,存储在寄存器711中的预测缩放因子可以被输入到量化器720。也就是说,当第一样本参数被输入时,与每个第一样本参数相对应的样本缩放因子可以通过缩放因子估计器710的一系列计算被顺序地输入到量化器720。在所有样本参数都被输入之后,可以计算预测缩放因子并将其存储在寄存器711中,并且可以将预测缩放因子输入到量化器720。

[0094] 量化器720可以通过乘法器MPR将预测缩放因子乘以每个输入数据0\_1、0\_2、...、和0\_N。随后,量化器720可以对被乘值进行舍入,并且可以通过限幅计算来输出最终量化的输出数据Q\_1、Q\_2、...、和Q\_N。量化器720中的舍入计算单元可以以整数的形式指示输出数据。

[0095] 限幅计算单元的计算可以被表达为[式9]。

[0096] [式9]

$$[0097] \quad \text{clip}(x, l, u) = \begin{cases} l & (\text{如果 } x < l) \\ x & (\text{如果 } l \leq x < u) \\ u & (\text{如果 } u < x) \end{cases}$$

[0098] 在[式9]中,x是输入数据值,l是待表达的量化的数据的最小值,并且u是待表达的量化的数据的最大值。量化器720中的限幅计算单元可以调整数据值,使得输出数据可以介于最大值与最小值之间。

[0099] 根据上述实施例,可以省略用于在深度学习推理期间的量化操作中针对所有输入估计缩放因子的计算,因此,总计算量可以减少。另外,用于基于量化的具有高准确度的人工神经网络来实现人工神经网络的计算资源可以减少,并且人工神经网络的应用范围可以扩展。

[0100] 在本文中,缩放因子估计器710和量化器710可以各自是包括以下一者或更多者的模拟电路和/或数字电路:逻辑门、集成电路、微处理器、微控制器、存储器电路、无源电子部件、有源电子部件、光部件等,并且可以被配置为运行软件和/或固件以执行上述对应的功

能或操作。

[0101] 图8是图示了根据示例实施例的计算系统2000的框图。

[0102] 在一些实施例中,图1的量化系统100和/或图5的量化设备500可以被实现在图8的计算系统2000中。如图8所示,计算系统2000可以包括系统存储器2100、处理器2300、存储装置2500、输入/输出(I/O)装置2700和通信连接2900。例如,包括在计算系统2000中的部件可以经由总线彼此通信地连接。

[0103] 系统存储器2100可以包括程序2120。程序2120可以使得处理器2300能够执行根据示例实施例的人工神经网络的量化。例如,程序2120可以包括能由处理器2300运行的多个命令,并且包括在程序2120中的多个命令可以由处理器2300运行,使得可以执行人工神经网络的量化。作为非限制性示例的系统存储器2100可以包括诸如静态随机存取存储器(SRAM)和动态随机存取存储器(DRAM)的易失性存储器,或诸如闪存的非易失性存储器等。

[0104] 处理器2300可以包括用于运行任意命令集的至少一个核(例如,英特尔架构-32(IA-32)、64位扩展IA-32、x86-64、PowerPC、Sparc、MIPS、ARM、IA-64等)。处理器2300可以运行存储在系统存储器2100中的命令并且可以运行程序2120,使得可以执行人工神经网络的量化。

[0105] 即使向计算系统2000供应的电力被阻断,存储装置2500也不会丢失所存储的数据。例如,存储装置2500可以包括诸如电可擦除可编程只读存储器(EEPROM)、闪存、相变随机存取存储器(PRAM)、电阻随机存取存储器(RRAM)、纳米浮栅存储器(NFGM)、聚合物随机存取存储器(PoRAM)、磁性随机存取存储器(MRAM)、铁电随机存取存储器(FRAM)或诸如磁带、光盘、磁盘的存储介质等的非易失性存储器。在一些实施例中,存储装置2500可以从计算系统2000可拆卸。

[0106] 在一些实施例中,存储装置2500可以存储用于根据示例实施例的人工神经网络的量化的程序2120,并且在程序2120由处理器2300运行之前,程序2120或其至少一部分可以被从存储装置2500加载到系统存储器2100中。在一些实施例中,存储装置2500可以存储用程序语言编写的文件,并且由编译器等生成的程序2120或其至少一部分可以被从文件加载到系统存储器2100中。

[0107] 在一些实施例中,存储装置2500可以存储将由处理器2300处理的数据和/或由处理器2300处理的数据。例如,存储装置2500可以存储根据上述量化方法量化的激活或数据,可以存储图1的量化的输出数据集OUT,或者可以存储在量化中生成的偏置或数据。

[0108] 输入/输出装置2700可以包括诸如键盘、指点装置等的输入装置,并且可以包括诸如显示装置、打印机等的输出装置。例如,用户可以触发由处理器2300运行程序2120,可以输入图2的输入数据(例如, $I_1$ 、 $I_2$ ),或者可以通过输入/输出装置2700来检查图1的量化的输出数据集OUT和/或错误消息。

[0109] 通信连接2900可以提供关于计算系统2000外部的网络的访问。例如,网络可以包括多个计算系统和通信链路,并且通信链路可以包括有线链路、光链路、无线链路或具有其他任意格式的链路。

[0110] 图9是图示了根据示例实施例的便携式计算系统的框图。

[0111] 在一些实施例中,根据示例实施例的量化的人工神经网络可以被实现在便携式计算装置3000中。作为非限制性示例的便携式计算装置3000可以是通过电池或自备电源来供

应电力的任何便携式电子装置,诸如移动电话、平板个人计算机(PC)、可穿戴装置和物联网(IoT)装置。

[0112] 如图9所示,便携式计算装置3000可以包括存储子系统3100、输入/输出(I/O)装置3300、处理单元3500和网络接口3700。存储子系统3100、输入/输出装置3300、处理单元3500和网络接口3700可以经由总线3900彼此通信。在一些实施例中,存储子系统3100、输入/输出装置3300、处理单元3500和网络接口3700中的至少两者可以作为片上系统(SoC)被包括在一个封装件中。

[0113] 存储子系统3100可以包括随机存取存储器(RAM) 3120和存储装置3140。RAM 3120和/或存储装置3140可以存储将由处理单元3500运行的命令和将由处理单元3500处理的数据。例如,RAM 3120和/或存储装置3140可以存储诸如人工神经网络的信号、权重、偏置的变量,并且也可以存储人工神经网络的人工神经元(或计算节点)的参数。在一些实施例中,存储装置3140可以包括非易失性存储器。

[0114] 处理单元3500可以包括CPU 3520、GPU 3540、数字信号处理器(DSP) 3560和NPU 3580。与在图9中不同,在一些实施例中,处理单元3500可以仅包括CPU 3520、GPU 3540、DSP 3560和NPU 3580的至少部分。

[0115] CPU 3520可以控制便携式计算装置3000的整体操作,并且它本身执行特定工作,或者可以例如响应于通过输入/输出装置3300接收到的外部输入来指示处理单元3500的其他部件执行特定工作。GPU 3540可以生成将通过包括在输入/输出装置3300中的显示设备输出的图像的数据,或者可以对从包括在输入/输出装置3300中的相机接收到的数据进行编码。DSP 3560可以处理数字信号,例如,从网络接口3700提供的数字信号,使得可以生成有效数据。

[0116] 作为用于人工神经网络的执行硬件的NPU 3580可以包括与构成人工神经网络的人工神经元的至少部分相对应的多个计算节点,并且多个计算节点的至少部分可以以并行方式处理信号。根据示例实施例,量化的人工神经网络(诸如深度神经网络)具有高准确度以及低计算复杂度,因此可以被容易地实现在图9的便携式计算装置3000中,并且量化的人工神经网络具有快的处理速度,并且可以由例如简单且小的NPU 3580实现。

[0117] 输入/输出装置3300可以包括诸如触摸输入装置、声音输入装置、相机等的输入装置以及诸如显示装置、声音输出装置等的输出装置。网络接口3700可以向便携式计算装置3000提供关于诸如长期演进(LTE)、第五代(5G)等的移动通信网络的访问,并且也可以提供关于诸如无线保真(WiFi)的局域网的访问。

[0118] 虽然已经参考本发明构思的实施例具体示出和描述了本发明构思,但是应理解,在不脱离所附权利要求及其等同形式的精神和范围的情况下,可以在其中做出形式和细节上的各种改变。

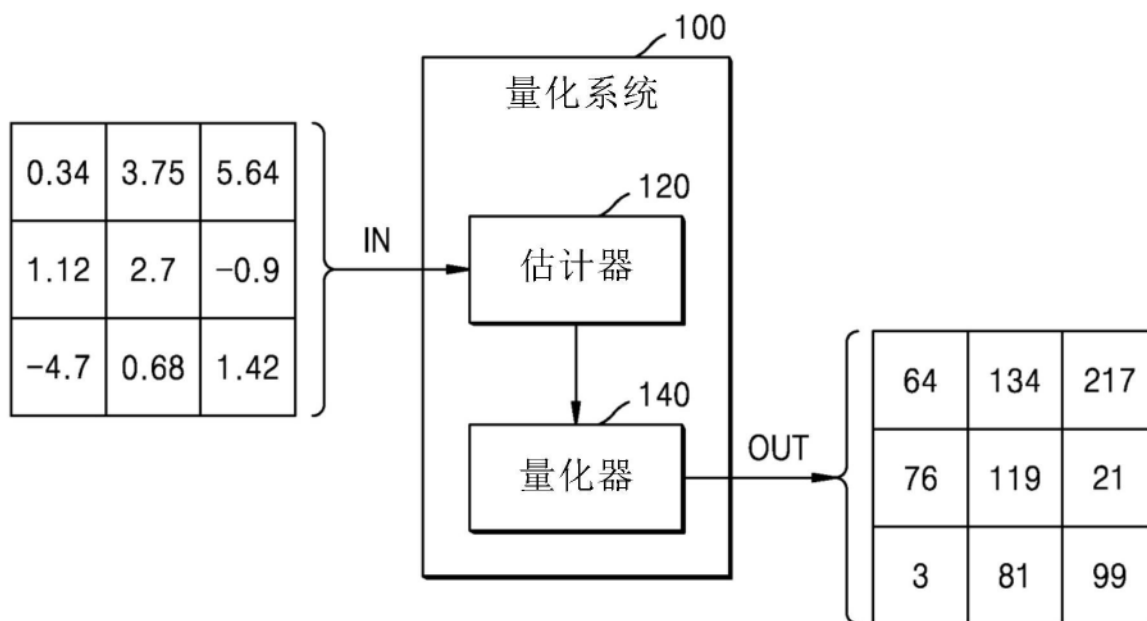


图1

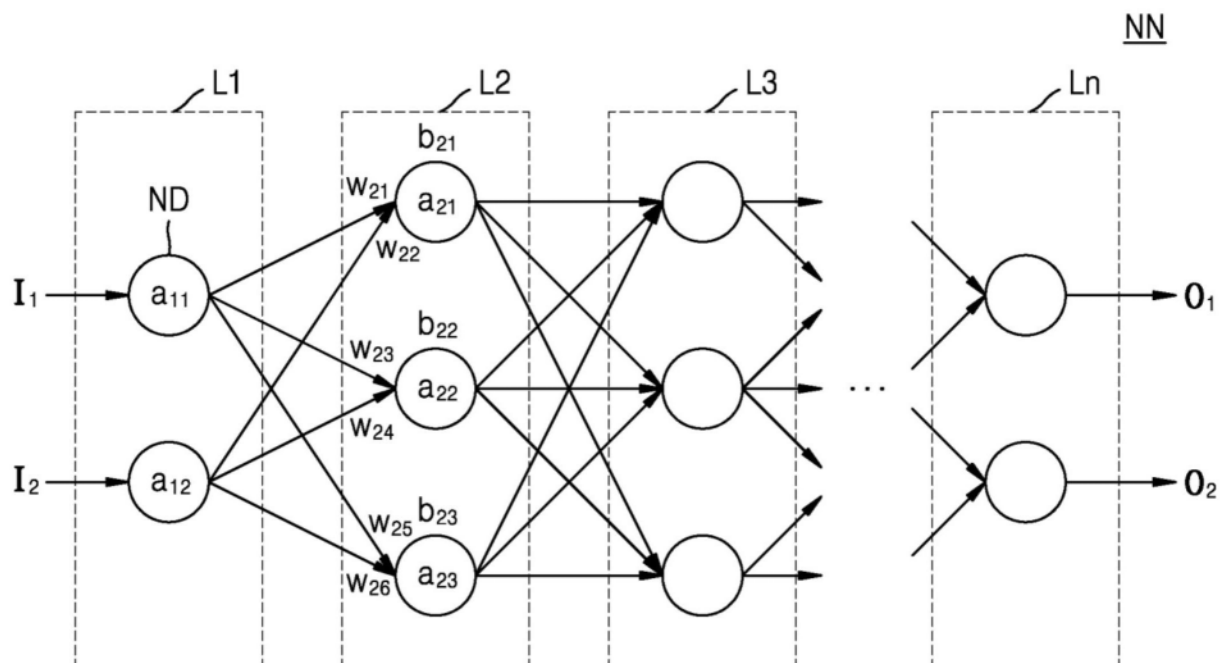


图2

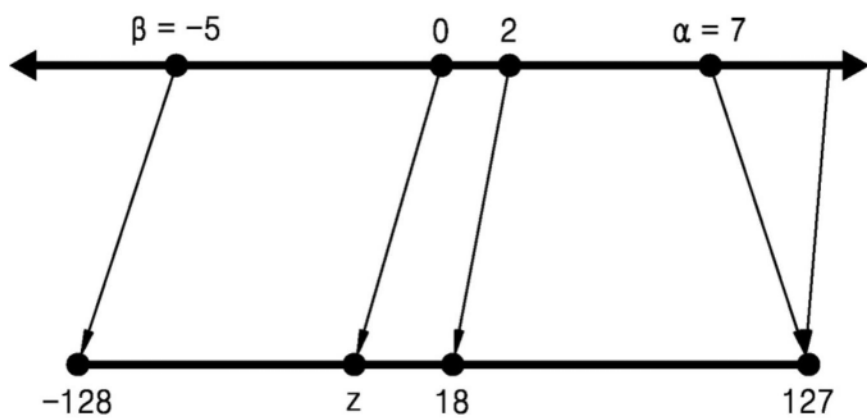


图3A

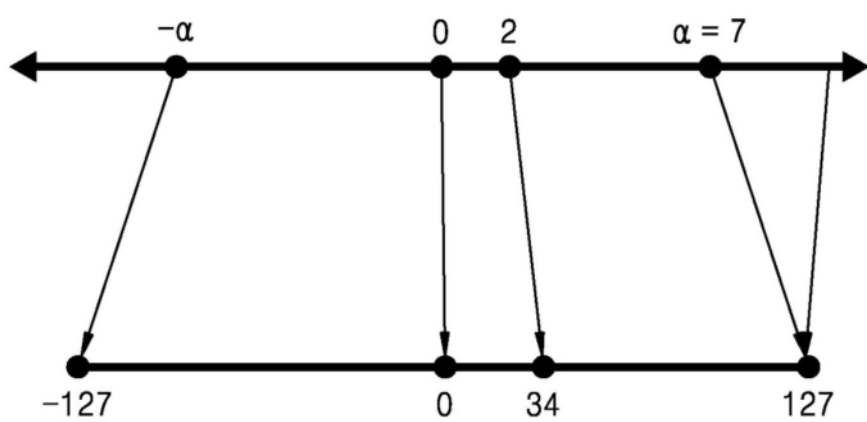


图3B

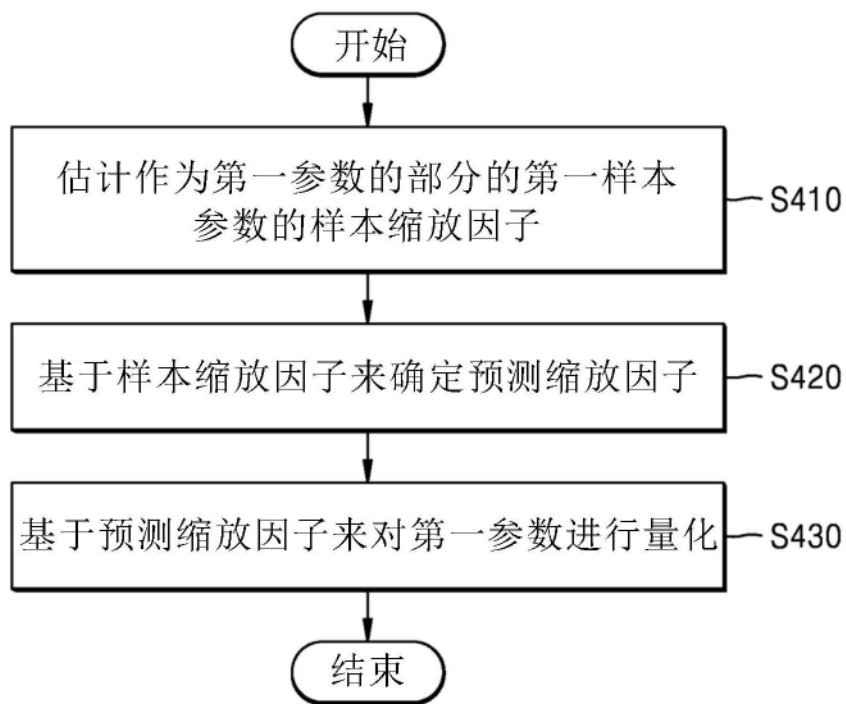


图4

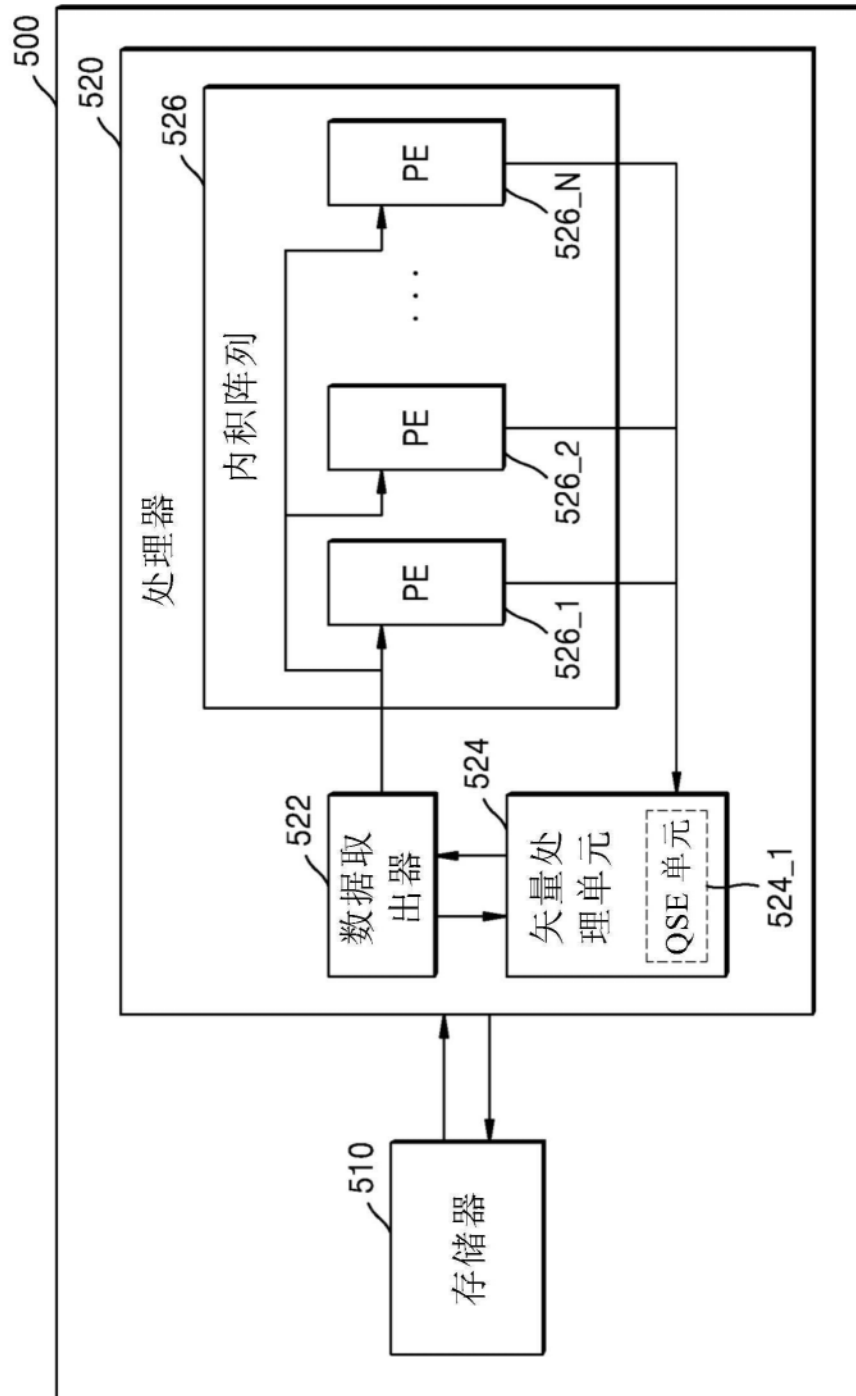


图5

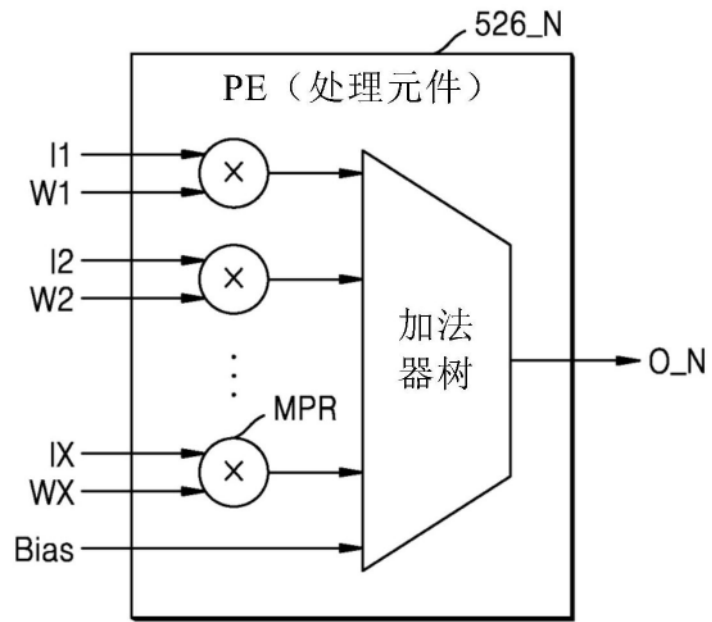


图6

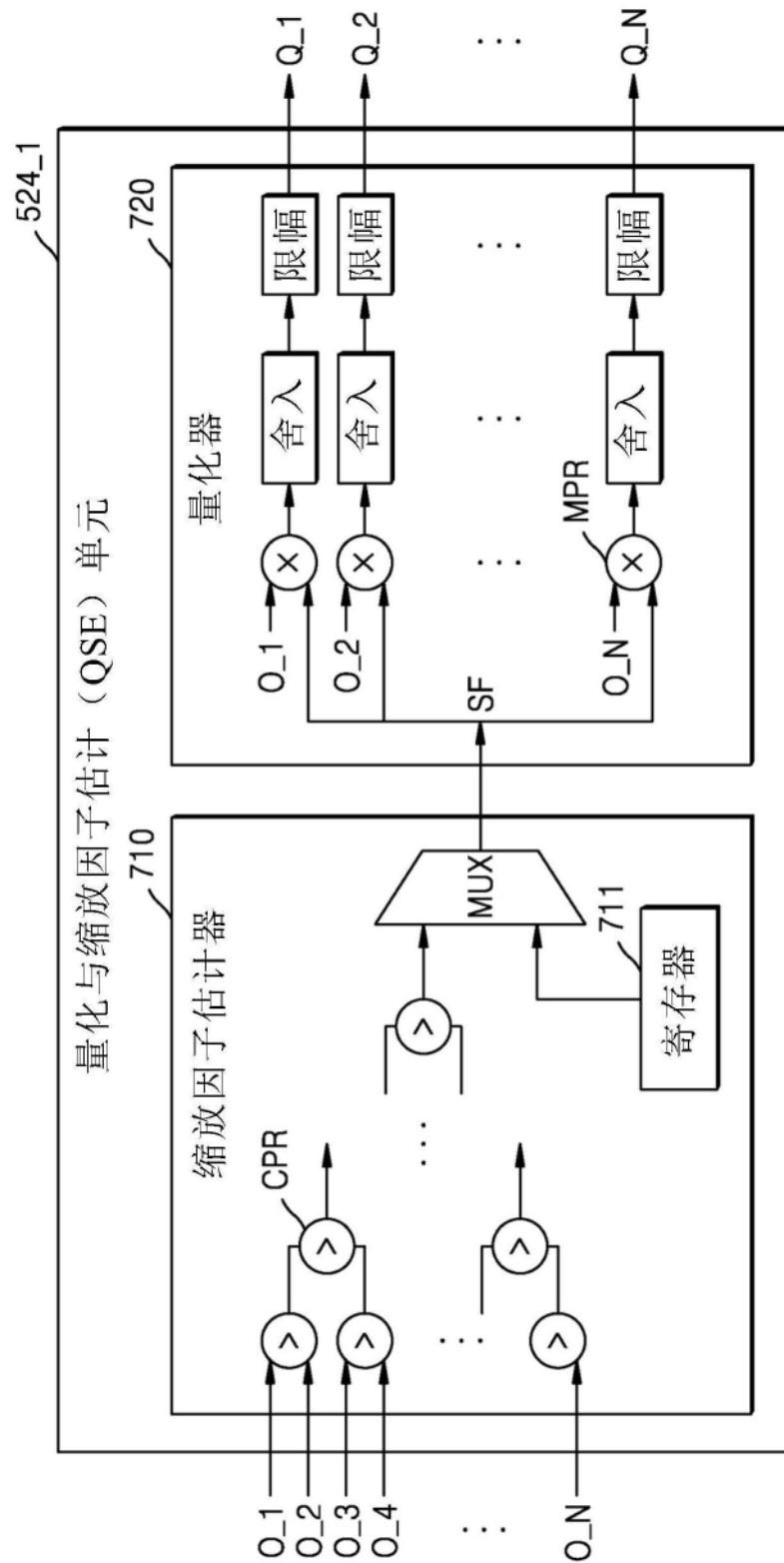


图7

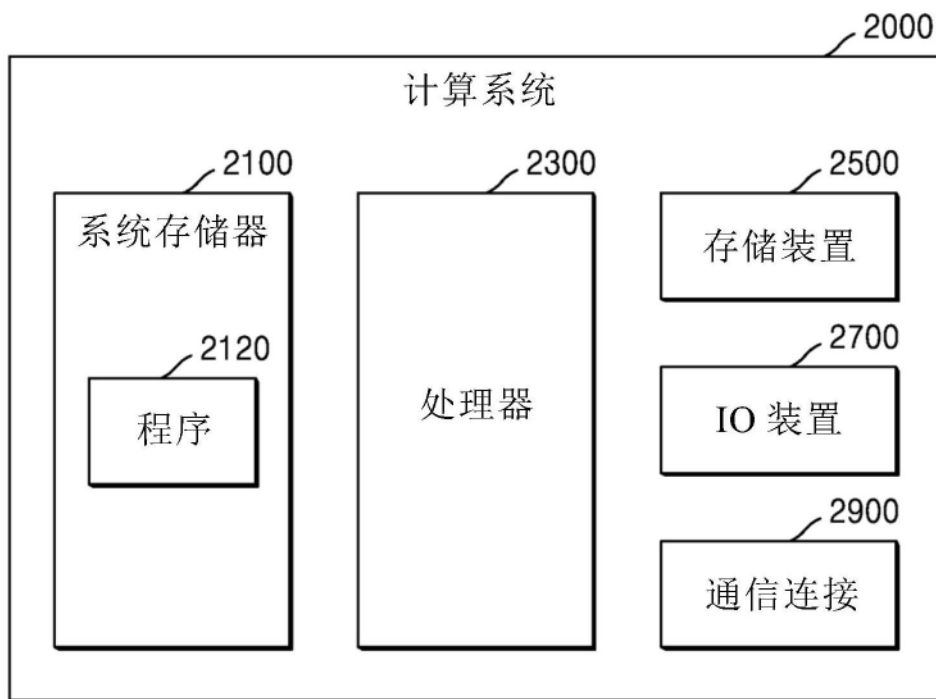


图8

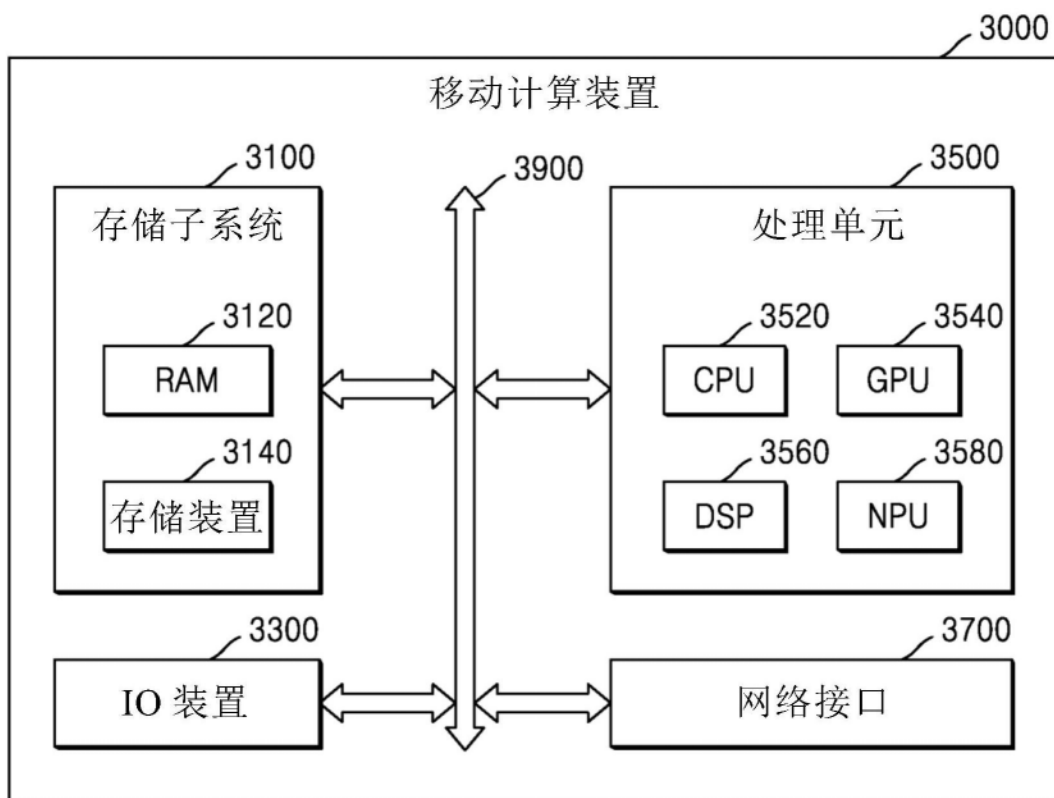


图9