(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 120297332 A
(43)申请公布日 2025.07.11

(21)申请号 202510033789.9

(22)申请日 2025.01.09

(30)优先权数据
10-2024-0004355 2024.01.10 KR
10-2024-0061261 2024.05.09 KR

(71)申请人 三星电子株式会社
地址 韩国京畿道水原市
申请人 高丽大学校产学协力团

(72)发明人 范山玉 徐泰源

(74)专利代理机构 北京铭硕知识产权代理有限公司 11286
专利代理师 黄晓燕 张川绪

(51)Int.Cl.
G06N 3/0464(2023.01)
G06N 3/0495(2023.01)
G06N 3/08(2023.01)
G06N 5/04(2023.01)

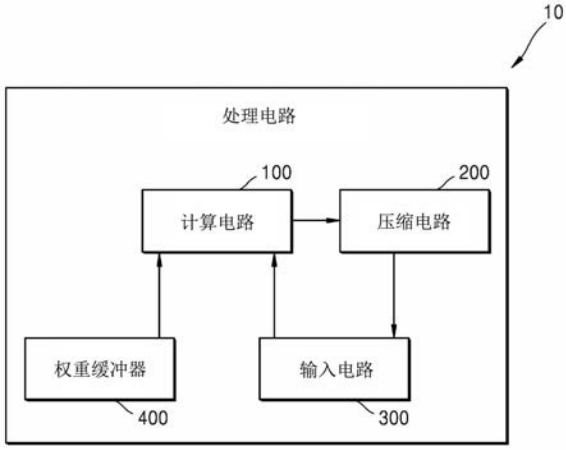
权利要求书3页 说明书15页 附图10页

(54)发明名称

处理电路、操作处理电路的方法以及用于人工神经网络的系统

(57)摘要

提供了处理电路、操作处理电路的方法以及用于人工神经网络的系统，所述处理电路被配置为基于人工神经网络中的第二层的步幅通过使用步幅感知压缩稀疏行(SCSR)算法从第一层的输出数据提取非零值来将第一层的输出数据压缩为第二压缩输入数据，并且输出第二压缩输入数据作为第二层的输入数据。



CN 120297332 A

1.一种处理电路,包括:

计算电路,被配置为生成人工神经网络中的第一层的输出数据,生成第一层的输出数据的处理包括基于第一层的输入数据和第一层的权重数据来执行卷积运算;以及

压缩电路,被配置为将第一层的输出数据压缩为第二压缩输入数据,并且输出第二压缩输入数据作为第二层的输入数据,将第一层的输出数据压缩为第二压缩输入数据的处理包括基于人工神经网络中的第二层的步幅从第一层的输出数据提取非零值,

其中,从第一层的输出数据提取非零值的处理包括:使用步幅感知压缩稀疏行算法,并且

其中,第二层是第一层的后续层。

2.根据权利要求1所述的处理电路,其中,压缩电路被配置为:对第一层的输出数据进行压缩,使得第二压缩输入数据的条数与第二层的步幅相同。

3.根据权利要求2所述的处理电路,其中,第一层的输出数据以包括多个元素的矩阵形式形成,并且

其中,第二压缩输入数据中的每个由所述多个元素之中的在所述多个元素的列索引除以第二层的步幅时具有相同余数值的元素组成。

4.根据权利要求1所述的处理电路,其中,第一层的输入数据包括与第一层的步幅相同的条数的第一压缩输入数据,并且

其中,第一层的权重数据包括基于第一层的步幅通过使用步幅感知压缩稀疏行算法生成的与第一层的步幅相同条数的第一压缩权重数据。

5.根据权利要求4所述的处理电路,其中,第一压缩权重数据在训练人工神经网络的操作中被生成。

6.根据权利要求4所述的处理电路,还包括:

输入电路,被配置为存储第一压缩输入数据和第二压缩输入数据;以及

权重缓冲器,被配置为存储第一压缩权重数据,

其中,计算电路包括以矩阵形式布置的多个处理元件,

其中,所述多个处理元件被配置为使得:具有相同行索引的处理元件被配置为从输入电路接收相同的压缩输入数据,并且具有相同列索引的处理元件被配置为从权重缓冲器接收相同的压缩权重数据。

7.根据权利要求6所述的处理电路,其中,所述多个处理元件各自包括:

乘法累加电路,被配置为通过基于从输入电路接收的压缩输入数据和从权重缓冲器接收的压缩权重数据执行乘法和累加来生成输出数据;

地址计算电路,被配置为基于从输入电路接收的压缩输入数据的地址信号来计算输出数据的地址;

输出缓冲器,被配置为存储输出数据的地址和输出数据;以及

处理元件控制器,被配置为控制输入电路将压缩输入数据的有效元素发送到乘法累加电路,并且控制权重缓冲器将压缩权重数据的有效元素发送到乘法累加电路。

8.根据权利要求7所述的处理电路,其中,

处理元件控制器还被配置为:基于小阵列、以及压缩权重数据的开始元素索引和结束元素索引使用卷积算法来计算输入计数器值和权重计数器值,控制输入电路基于输入计数

器值将压缩输入数据的元素之中的有效元素发送到乘法累加电路,并且控制权重缓冲器基于权重计数器值将压缩权重数据的元素之中的有效元素发送到乘法累加电路,

其中,所述小阵列表示用于选择输入计数器值的初始值的阵列,

压缩权重数据的开始元素索引表示压缩权重数据的有效元素之中的具有最小的行索引的元素的索引,并且

压缩权重数据的结束元素索引表示压缩权重数据的有效元素之中的具有最大的行索引的元素的索引。

9.根据权利要求8所述的处理电路,还包括:

主控制器,被配置为:

基于压缩输入数据来生成所述小阵列、压缩权重数据的开始元素索引和压缩权重数据的结束元素索引,并且

将所述小阵列、压缩权重数据的开始元素索引和压缩权重数据的结束元素索引发送给处理元件控制器。

10.根据权利要求1所述的处理电路,其中,第二压缩输入数据在对人工神经网络进行推断的操作中被生成。

11.一种操作处理电路的方法,所述方法包括:

通过处理电路的计算电路通过基于人工神经网络中的第一层的输入数据和第一层的权重数据执行卷积运算来生成第一层的输出数据;

通过处理电路的压缩电路从第一层的输出数据提取非零值;以及

通过压缩电路基于提取的非零值和人工神经网络中的第二层的步幅,将第一层的输出数据压缩为第二压缩输入数据,

其中,第二层是第一层的后续层。

12.根据权利要求11所述的方法,其中,第二压缩输入数据的条数与第二层的步幅相同。

13.根据权利要求11所述的方法,其中,第一层的输出数据以包括多个元素的矩阵形式形成,并且

其中,将第一层的输出数据压缩为第二压缩输入数据的步骤包括:由所述多个元素之中的在所述多个元素的列索引除以第二层的步幅时具有相同余数值的元素组成第二压缩输入数据。

14.根据权利要求11所述的方法,还包括:

通过压缩电路从先前层的输出数据提取非零值,第一层在所述先前层之后;以及

通过压缩电路基于从所述先前层的输出数据提取的非零值和第一层的步幅,将所述先前层的输出数据压缩为第一压缩输入数据,

其中,第一层的输入数据包括第一压缩输入数据。

15.根据权利要求14所述的方法,还包括:

训练人工神经网络,

其中,第一层的权重数据在人工神经网络的训练中基于第一层的步幅而被生成,并且包括与第一层的步幅相同的条数的第一压缩权重数据。

16.根据权利要求15所述的方法,其中,生成第一层的输出数据的步骤包括:

基于第一压缩输入数据来确定输入计数器值和权重计数器值;

接收第一压缩权重数据的有效元素和第一压缩输入数据的有效元素,第一压缩权重数据的有效元素基于权重计数器值,并且第一压缩输入数据的有效元素基于输入计数器值;以及

通过基于第一压缩输入数据的有效元素和第一压缩权重数据的有效元素执行卷积运算来生成第一层的输出数据。

17.一种用于人工神经网络的系统,包括:

至少一个处理器;以及

非暂时性存储介质,存储指令,所述指令被配置为当被所述至少一个处理器执行时使得所述至少一个处理器执行对人工神经网络中的多个层的输出数据进行压缩的方法,

其中,对人工神经网络中的所述多个层的输出数据进行压缩的方法包括:

生成人工神经网络中的第一层的输出数据,生成人工神经网络中的第一层的输出数据的步骤包括基于人工神经网络中的第一层的输入数据和第一层的权重数据来执行卷积运算,

从第一层的输出数据提取非零值,以及

基于提取的非零值和人工神经网络中的第二层的步幅,将第一层的输出数据压缩为第二压缩输入数据,

其中,第二层是第一层的后续层。

18.根据权利要求17所述的系统,其中,第二压缩输入数据的条数与第二层的步幅相同。

19.根据权利要求17所述的系统,其中,第一层的输出数据以包括多个元素的矩阵形式形成,并且

其中,将第一层的输出数据压缩为第二压缩输入数据的步骤包括:由所述多个元素之中的在所述多个元素的列索引除以第二层的步幅时具有相同余数值的元素组成第二压缩输入数据。

20.根据权利要求17所述的系统,其中,所述指令被配置为当被所述至少一个处理器执行时使得所述至少一个处理器还执行:

从第一层的先前层的输出数据提取非零值;并且

基于从所述先前层的输出数据提取的非零值和第一层的步幅,将所述先前层的输出数据压缩为第一压缩输入数据,

其中,第一层的输入数据包括第一压缩输入数据,并且

其中,人工神经网络被用于处理多媒体数据,并且多媒体数据包括语音数据、图像数据、视频数据和文本数据中的至少一者。

处理电路、操作处理电路的方法以及用于人工神经网络的系统

[0001] 本申请基于并要求于2024年1月10日在韩国知识产权局提交的第10-2024-0004355号韩国专利申请和于2024年5月9日在韩国知识产权局提交的第10-2024-0061261号韩国专利申请的优先权,所述韩国专利申请的公开通过引用全部包含于此。

技术领域

[0002] 发明构思涉及人工神经网络,并且更具体地,涉及用于人工神经网络的处理电路、操作处理电路的方法以及包括处理电路的系统。

背景技术

[0003] 人工神经网络可表示计算装置或由该计算装置执行以实现人工神经元(或神经元模型)的互连集合的方法。人工神经元可通过对输入数据执行简单的运算来生成输出数据,其中,输出数据可被传递给另外的人工神经元。作为人工神经网络的示例的深度神经网络或深度学习可具有多层结构。

[0004] 由于深度学习推断计算需要大量的计算,因此在有限的环境(诸如,移动环境和/或需要高速处理的环境)中,人工神经网络的有用性可能被限制。因此,可能需要用于高效地对模型的每个层的输入进行压缩的方法。

发明内容

[0005] 发明构思提供能够高效地对人工神经网络中的每个层的输入进行压缩的处理装置、操作处理装置的方法以及包括处理装置的系统。

[0006] 根据发明构思的一个方面,提供了一种处理电路,所述处理电路包括:计算电路,被配置为生成人工神经网络(ANN)中的第一层的输出数据,生成第一层的输出数据的处理包括基于第一层的输入数据和第一层的权重数据来执行卷积运算;以及压缩电路,被配置为将第一层的输出数据压缩为第二压缩输入数据,并且输出第二压缩输入数据作为第二层的输入数据,对第一层的输出数据进行压缩的处理包括基于人工神经网络中的第二层的步幅从第一层的输出数据提取非零值,其中,从第一层的输出数据提取非零值的处理包括:使用步幅感知压缩稀疏行(SCSR)算法,并且其中,第二层是第一层的后续层。

[0007] 根据发明构思的另一方面,提供了一种操作处理电路的方法,所述方法包括:通过基于人工神经网络中的第一层的输入数据和第一层的权重数据执行卷积运算来生成第一层的输出数据;从第一层的输出数据提取非零值;以及基于提取的非零值和人工神经网络中的第二层的步幅,将第一层的输出数据压缩为第二压缩输入数据,其中,第二层是第一层的后续层。

[0008] 根据发明构思的另一方面,提供了一种系统,所述系统包括:至少一个处理器;以及非暂时性存储介质,存储指令,所述指令被配置为当被所述至少一个处理器执行时使得所述至少一个处理器执行对人工神经网络中的多个层的输出数据进行压缩的方法,其中,

对人工神经网络中的所述多个层的输出数据进行压缩的方法包括:基于人工神经网络中的第一层的输入数据和第一层的权重数据来执行卷积运算并且生成第一层的输出数据,从第一层的输出数据提取非零值,以及基于人工神经网络中的第二层的步幅和提取的非零值来将第一层的输出数据压缩为第二压缩输入数据,其中,第二层是第一层的后续层。

附图说明

[0009] 从以下结合附图的详细描述将更清楚地理解实施例。

[0010] 图1是根据至少一个实施例的处理电路的框图。

[0011] 图2是根据至少一个实施例的操作处理电路的方法的流程图。

[0012] 图3A至图3E是示出根据至少一个实施例的操作处理电路的方法和根据对比实施例的操作处理电路的方法的示图。

[0013] 图4A和图4B是示出根据至少一个实施例的处理电路的对人工神经网络中的每个层的输入数据和权重数据进行压缩的处理的示图。

[0014] 图5是根据至少一个实施例的处理电路的框图。

[0015] 图6是根据至少一个实施例的处理元件(PE)的框图。

[0016] 图7是根据至少一个实施例的处理电路的卷积运算的流程图。

[0017] 图8是示出根据至少一个实施例的处理电路的卷积运算的示图。

[0018] 图9示出用于将根据至少一个实施例的处理电路与根据对比实施例的处理电路在处理电路的每个组件的功耗和功耗百分比方面进行比较的图。

[0019] 图10是根据至少一个实施例的计算系统的框图。

[0020] 图11是根据至少一个实施例的便携式计算装置的框图。

具体实施方式

[0021] 在下文中,参照附图详细描述实施例。

[0022] 图1是根据至少一个实施例的处理电路的框图。

[0023] 以下描述的功能和/或实现所述功能的功能元件可由处理电路系统(诸如,硬件、软件、或者硬件和软件的组合)来实现或支持。例如,处理电路系统可包括但不限于中央处理器(CPU)、应用处理器(AP)、算术逻辑单元(ALU)、图形处理器(GPU)、数字信号处理器、微计算机、现场可编程门阵列(FPGA)、片上系统(SoC)、可编程逻辑单元、微处理器或专用集成电路(ASIC)等。例如,以下描述的各种功能可由人工智能技术或者一个或多个计算机程序来实现或支持,一个或多个计算机程序中的每个由计算机可读程序代码组成并且在计算机可读介质上实现。术语“应用”和“程序”表示适合于在合适的计算机可读程序代码中实现的一个或多个计算机程序、软件组件、指令集、进程、函数、对象、类、实例、相关联的数据或它们的部分。术语“计算机可读程序代码”包括所有类型的计算机代码(包括源代码、目标代码和可执行代码)。术语“计算机可读介质”包括可由计算机访问的任何类型的介质(诸如,只读存储器(ROM)、随机存取存储器(RAM)、硬盘驱动器、光盘(CD)、数字视频盘(DVD)或任何其他类型的存储器)。“非暂时性”计算机可读介质排除发送瞬时电信号或其他信号的有线通信链路、无线通信链路、光学通信链路或其他通信链路。非暂时性计算机可读介质包括数据可永久地被存储在其中的介质、以及数据可被存储在其中并随后被重写的介质(诸如,可重

写光盘或可擦除存储器装置)。

[0024] 在以下描述的一些实施例中,基于硬件的方法被示出为示例。然而,由于实施例包括使用硬件和软件两者的技术,因此实施例不排除基于软件的(例如,基于软件启用的)方法。

[0025] 人工神经网络(ANN)可表示由构成动物大脑的生物神经网络启发的计算系统。在一个实施例中,ANN可被用于处理多媒体数据,并且多媒体数据可包括语音数据、图像数据、视频数据和文本数据中的至少一者。与根据预定义的条件执行任务的常规算法(诸如,基于规则的编程)不同,ANN可通过考虑多个样本(或示例)来学习执行任务。ANN可具有其中人工神经元(或神经元)被连接的结构,其中,神经元之间的连接可被称为突触。神经元可对接收的信号进行处理,并且通过突触将处理后的信号发送到其他神经元。神经元的输出可被称为激活(activation)。与动物大脑不同,ANN的神经元和/或突触可具有可变的权重,并且根据权重,由神经元处理的信号的影响可增大或减小。具体地,与各个神经元相关联的权重可被称为偏置(bias)。

[0026] ANN可具有分层的结构。例如,ANN可以是(和/或包括)深度神经网络(DNN)或深度学习架构;并且深度神经网络(DNN)或深度学习架构可具有多层的结构,并且特定层的输出可变为后续层的输入。例如,DNN可包括但不限于卷积神经网络(CNN)、循环神经网络(RNN)、深度信念网络、受限玻尔兹曼机等。在该多层的结构中,每个层可基于多个样本而被训练。ANN(诸如,DNN)可由分别与人工神经元对应的多个处理节点来实现。为了获得良好的结果(诸如,具有高准确度的结果),可能需要高计算复杂度和许多计算资源。

[0027] 参照图1,根据至少一个实施例的处理电路10包括计算电路100、压缩电路200、输入电路300和权重缓冲器400。计算电路100可被配置为从输入电路300接收ANN中的每个层的输入数据,并且从权重缓冲器400接收ANN中的每个层的权重数据。计算电路100可被配置为通过基于接收的输入数据和权重数据执行卷积运算来生成ANN中的每个层的输出数据。卷积运算可表示对有效输入对(例如,基于步幅(stride,又称为步长)执行有效乘法的“输入数据的元素和权重数据的元素”的对)执行乘法并对乘法的结果执行累加的运算。步幅可表示当执行卷积运算时权重数据在行方向或列方向上的移动的量。输入数据和权重数据可以以矩阵形式(例如,稀疏矩阵)形成。稀疏矩阵可表示其中大多数元素的值为0的矩阵。当对不影响输出数据的具有值0的元素执行卷积运算时,处理电路10可通过从输入数据或权重数据提取不具有值0的元素(即,具有非零值的元素)来对输入数据或权重数据进行压缩。当对压缩输入数据(或称为压缩后的输入数据)或压缩权重数据(或称为压缩后的权重数据)执行卷积运算时,计算复杂度可被降低,但是也可对执行无效乘法的输入对执行乘法。因此,搜索有效输入对的操作可能是必要的。搜索有效输入对的操作可增大处理电路10的功耗。

[0028] 为了避免在对输入数据或权重数据进行压缩的同时执行搜索有效输入对的操作,根据发明构思的压缩电路200可被配置为基于ANN中的每个层的步幅,通过步幅感知压缩稀疏行(stride-aware compressed sparse row,SCSR)算法来对输入数据和/或权重数据进行压缩。SCSR算法可表示:从输入数据的元素或权重数据的元素提取非零值,并且通过将以下的等式1应用于具有非零值的元素来生成压缩输入数据和/或压缩权重数据。

[0029] [等式1]

$$t_f = c_f^o \ mod \ s$$

$$t_w = c_w^o \ mod \ s$$

[0030] 在等式1中，t_f表示压缩输入数据的索引；c_f^o表示输入数据的列索引；t_w表示压缩权重数据的索引；c_w^o表示权重数据的列索引；S表示步幅；并且mod表示用于计算余数的数学符号。以下可参照图3B、图4A和图4B描述通过应用等式1来生成压缩输入数据或压缩权重数据的特定实施例。

[0031] 在一些实施例中，压缩电路200可从计算电路100接收ANN中的特定层的输出数据，并且可基于后续层的步幅使用SCSR算法来将特定层的输出数据压缩为压缩输入数据。

[0032] 例如，计算电路100可被配置为通过基于第一层的输入数据和第一层的权重数据执行卷积运算来生成第一层的输出数据，并且可将第一层的输出数据发送到压缩电路200。压缩电路200可基于第二层的步幅通过使用SCSR算法从第一层的输出数据提取非零值，来将第一层的输出数据压缩为第二压缩输入数据，第二层是第一层的后续层。第二压缩输入数据可被用作第二层的输入数据。

[0033] 例如，压缩电路200可将第一层的输出数据压缩为与第二层的步幅相同的条数的第二压缩输入数据。当第二层的步幅是2时，第二压缩输入数据的条数可以是2，并且当第二层的步幅是3时，第二压缩输入数据的条数可以是3。

[0034] 例如，压缩电路200可使得第二压缩输入数据中的每个由第一层的输出数据的多个元素之中的“在多个元素的列索引除以第二层的步幅时具有相同的余数值的元素”组成。当第二层的步幅是2时，第二压缩输入数据的特定压缩输入数据可由第一层的输出数据的多个元素之中的列索引为偶数的元素组成，并且第二压缩输入数据的其他压缩输入数据可由第一层的输出数据的多个元素之中的列索引为奇数的元素组成。

[0035] 在一些实施例中，压缩电路200可使用SCSR算法将特定层的输出数据压缩为压缩输入数据并将压缩输入数据发送到输入电路300，并且计算电路100可从输入电路300接收压缩输入数据。权重缓冲器400可存储使用SCSR算法被压缩的压缩权重数据，并且计算电路100可从权重缓冲器400接收压缩权重数据。计算电路100可通过基于接收的压缩输入数据和压缩权重数据执行卷积运算来生成输出数据。

[0036] 例如，压缩电路200可被配置为通过使用SCSR算法将第一层的输出数据压缩为第二压缩输入数据，并且将第二压缩输入数据发送到输入电路300。权重缓冲器400可存储第二压缩权重数据，第二层的权重数据通过使用SCSR算法被压缩为第二压缩权重数据。计算电路100可从输入电路300接收第二压缩输入数据，并且从权重缓冲器400接收第二压缩权重数据。计算电路100可通过基于第二压缩输入数据和第二压缩权重数据执行卷积运算来生成第二层的输出数据。当第二层的步幅是2时，第二压缩输入数据的条数可以是2，并且第二压缩权重数据的条数可以是2。

[0037] 例如，压缩电路200可被配置为通过使用SCSR算法输出与第一层的步幅相同的条数的第一压缩输入数据，并且输入电路300可接收并存储第一压缩输入数据。第一层的权重数据可包括使用SCSR算法生成的与第一层的步幅相同的条数的第一压缩权重数据，并且可被存储在权重缓冲器400中。计算电路100可通过基于从输入电路300接收的第一压缩输入数据和从权重缓冲器400接收的第一压缩权重数据执行卷积运算来生成第一层的输出数

据,其中,以下可参照图5至图8描述卷积运算的特定实施例。

[0038] 输入电路300可被配置为将通过使用SCSR算法被压缩的压缩输入数据存储为ANN中的每个层的输入数据。权重缓冲器400可将通过使用SCSR算法被压缩的压缩权重数据存储为ANN中的每个层的权重数据。作为非限制性示例的输入电路300或权重缓冲器400可包括易失性存储器(诸如,静态随机存取存储器(SRAM)和动态随机存取存储器(DRAM)),并且还可包括非易失性存储器(诸如,闪存)。

[0039] 在图1中,处理电路10被示出为包括权重缓冲器400,但是可不限于此。例如,权重缓冲器400可被配置为处理电路10外部的单独存储器。

[0040] 根据发明构思的压缩电路200可基于ANN中的每个层的步幅通过使用SCSR算法来对每个层的输出数据进行压缩,其中,压缩输出数据可被用作后续层的输入数据。当基于ANN中每个层的步幅使用SCSR算法时,有效输入对被压缩。因此,用于搜索有效输入对的单独电路可不被添加,或者存储空间可不被分配。因此,输入数据可在不执行搜索有效输入对的操作的情况下被压缩,从而减少处理电路10的功耗并且提高处理速度。

[0041] 图2是根据至少一个实施例的操作处理电路的方法的流程图。参照图2,操作处理电路的方法20可包括多个操作S210至S230。

[0042] 进一步参照图1,在操作S210中,处理电路10可通过执行卷积运算来生成第一层的输出数据。在一些实施例中,计算电路100可通过基于ANN中的第一层的输入数据和第一层的权重数据执行卷积运算来生成第一层的输出数据。以下可参照图5至图8描述卷积运算的特定实施例。

[0043] 在操作S220中,处理电路10可从输出数据提取非零值。在一些实施例中,压缩电路200可从计算电路100接收ANN中的第一层的输出数据,其中,第一层的输出数据可包括稀疏矩阵。压缩电路200可使用SCSR算法来提取第一层的输出数据的元素之中的不具有值0的元素(例如,具有非零值的元素)。例如,通过使用SCSR算法提取具有非零值的元素的操作可与通过使用压缩稀疏行(CSR)算法提取具有非零值的元素的操作相同。

[0044] 在操作S230中,处理电路10可基于第二层的步幅和提取的非零值将第一层的输出数据压缩为第二压缩输入数据,第二层是第一层的后续层。第二压缩输入数据可被用作第二层的输入数据。

[0045] 在一些实施例中,压缩电路200可将第一层的输出数据压缩为与第二层的步幅相同的条数的第二压缩输入数据。例如,当第二层的步幅是2时,第二压缩输入数据的条数可以是2,并且当第二层的步幅是3时,第二压缩输入数据的条数可以是3。

[0046] 在一些实施例中,压缩电路200可使得第二压缩输入数据中的每个由第一层的输出数据的多个元素之中的“在多个元素的列索引除以第二层的步幅时具有相同的余数值的元素”组成。例如,当第二层的步幅是2时,第二压缩输入数据的特定压缩输入数据可由第一层的输出数据的多个元素之中的列索引为偶数的元素组成,并且第二压缩输入数据的其他压缩输入数据可由第一层的输出数据的多个元素之中的列索引为奇数的元素组成。

[0047] 在一些实施例中,该方法还可包括:在操作S210之前,从第一层的先前层的输出数据提取非零值,并且基于从该先前层的输出数据提取的非零值和第一层的步幅将先前层的输出数据压缩为第一压缩输入数据。例如,第一层的输入数据可包括由压缩电路200使用SCSR算法输出的与第一层的步幅相同的条数的第一压缩输入数据。

[0048] 在一些实施例中,该方法还可包括:在操作S210之前训练ANN。ANN的训练可表示生成ANN中的每个层的权重数据。例如,第一层的权重数据可在ANN的训练中基于第一层的步幅而被生成。例如,在ANN的训练中,第一层的权重数据可包括基于第一层的步幅通过使用SCSR算法生成的与第一层的步幅相同的条数的第一压缩权重数据。

[0049] 图3A至图3E是示出根据至少一个实施例的操作处理电路的方法和根据对比实施例的操作处理电路的方法的示图。

[0050] 图3A是示出通过对未压缩的ANN中的特定层的一维输入数据(a)和一维权重数据(b)执行卷积运算来生成一维输出数据(c)的处理的图。可假设特定层的步幅是2。一维输入数据(a)可包括具有值0的元素(例如,f_1、f_4和f_6),并且一维权重数据(b)可包括具有值0的元素(例如,W_1)。

[0051] 当对不影响输出数据的具有值0的元素执行卷积运算时,图1中的处理电路10可通过从输入数据或权重数据提取不具有值0的元素(例如,具有非零值的元素)来对输入数据或权重数据进行压缩。

[0052] 图3B是示出根据至少一个实施例的通过使用SCSR算法对处理电路的ANN中的特定层的输入数据和权重数据进行压缩的操作的示图。进一步参照图1,压缩电路200可基于特定层的步幅通过使用SCSR算法从一维输入数据(a)提取非零值(例如,f_0、f_2、f_3、f_5和f_7),来将一维输入数据(a)压缩为两个一维压缩输入数据(a1和a2)。

[0053] 例如,一维输入数据(a)的多个元素可按照从第一元素f_0到第八元素f_7的顺序具有0到7的列索引。一维压缩输入数据(a1)可由一维输入数据(a)的多个元素之中的“具有非零值并且在通过使用等式1将多个元素的列索引除以2时具有余数值0”的元素(例如,f_0和f_2)组成,2是特定层的步幅。由于余数值是0,因此一维压缩输入数据(a1)的索引可以是0。一维压缩输入数据(a2)可由一维输入数据(a)的多个元素之中的“具有非零值并且在通过使用等式1将多个元素的列索引除以2时具有余数值1”的元素(例如,f_3、f_5和f_7)组成,2是特定层的步幅。由于余数值是1,因此一维压缩输入数据(a2)的索引可以是1。

[0054] 一维权重数据(b)可基于特定层的步幅通过使用SCSR算法被压缩为两个一维压缩权重数据(b1和b2)。例如,一维权重数据(b)的多个元素可按照从第一元素W_0到第四元素W_3的顺序具有0到3的列索引。一维压缩权重数据(b1)可由一维权重数据(b)的多个元素之中的“具有非零值并且在通过使用等式1将多个元素的列索引除以2时具有余数值0”的元素(例如,W_0和W_2)组成,2是特定层的步幅。由于余数值是0,因此一维压缩权重数据(b1)的索引可以是0。一维压缩权重数据(b2)可由一维权重数据(b)的多个元素之中的“具有非零值并且在通过使用等式1将多个元素的列索引除以2时具有余数值1”的元素(例如,W_3)组成,2是特定层的步幅。由于余数值是1,因此一维压缩权重数据(b2)的索引可以是1。

[0055] 图3C是示出根据对比实施例的通过使用CSR算法对处理电路的ANN中的特定层的输入数据和权重数据进行压缩的操作的示图。根据对比实施例的处理电路可通过经由使用CSR算法从一维输入数据(a)提取非零值(例如,f_0、f_2、f_3、f_5和f_7)来将一维输入数据(a)压缩为一维压缩输入数据(a3),并且可通过从一维权重数据(b)提取非零值(例如,W_0、W_2和W_3)来将一维权重数据(b)压缩为一维压缩权重数据(b3)。

[0056] 图3D是示出根据至少一个实施例的处理电路的压缩输入数据和压缩权重数据的卷积运算的示图,并且图3E是示出根据对比实施例的处理电路的压缩输入数据和压缩权重

数据的卷积运算的示图。图3D的图31可示出一维压缩输入数据(a1和a2)的元素和一维压缩权重数据(b1和b2)的元素随时间的卷积运算,并且图3E的图32可示出一维压缩输入数据(a3)的元素和一维压缩权重数据(b3)的元素随时间的卷积运算。

[0057] 当参照图31时,一维压缩输入数据(a1和a2)的元素和一维压缩权重数据(b1和b2)的元素的对是执行有效乘法的有效输入对。因此,通过对有效输入对执行乘法并且对乘法结果执行累加而生成的输出数据可与图3A的一维输出数据(c)相同。

[0058] 相比之下,参照图32,一维压缩输入数据(a3)的元素和一维压缩权重数据(b3)的元素的对可包括执行无效乘法的输入对。例如,由于特定层的步幅是2,因此一维压缩输入数据(a3)的元素f_3与一维压缩权重数据(b3)的元素W_0的乘法可能无效,一维压缩输入数据(a3)的元素f_3与一维压缩权重数据(b3)的元素W_2的乘法可能无效,并且一维压缩输入数据(a3)的元素f_5与一维压缩权重数据(b3)的元素W_2的乘法可能无效。因此,由于可对无效输入对执行乘法,因此输出数据可与图3A的一维输出数据(c)不同。

[0059] 换言之,根据对比实施例的处理装置甚至可对执行无效乘法的输入对执行乘法。因此,在不执行压缩的情况下生成的输出数据可与在执行压缩之后生成的输出数据不同。因此,根据对比实施例的处理装置可能需要搜索有效输入对的操作,并且用于搜索有效输入对的操作的功耗可增大。

[0060] 相反,由于处理装置基于特定层的步幅通过使用SCSR算法来对输入数据或权重数据进行压缩,因此根据至少一个实施例的处理装置可不执行搜索有效输入对的操作,并且可不添加用于搜索有效输入对的单独电路或者可不分配存储空间。因此,由于输入数据和/或权重数据可在不执行搜索有效输入对的操作的情况下被压缩,因此可在减少根据至少一个实施例的处理电路的功耗的同时提高处理速度。

[0061] 图4A和图4B是示出根据至少一个实施例的处理电路的对ANN中的特定层的输入数据和权重数据进行压缩的处理的示图。

[0062] 图4A是示出根据至少一个实施例的处理电路的基于特定层的步幅使用SCSR算法对ANN中的特定层的输入数据进行压缩的操作的示图。特定层的输入数据41可示出以16×16矩阵的形式形成的稀疏矩阵的第一行至第四行,并且特定层的步幅可以是2。

[0063] 在一些实施例中,进一步参照图1,压缩电路200可基于特定层的步幅通过使用SCSR算法从输入数据41提取非零值来将输入数据41压缩为两条压缩输入数据$T_f[0]$和$T_f[1]$。

[0064] 例如,压缩输入数据$T_f[0]$可由输入数据41的多个元素之中的“通过使用图1的等式1而具有非零值并且在多个元素的列索引|C_f|除以2时具有余数值0”的元素组成,2是特定层的步幅。由于余数值是0,因此压缩输入数据$T_f[0]$的索引可以是0。压缩输入数据$T_f[0]$可包括指示输入数据41的行地址的行指针R_f、指示输入数据41的列地址的列索引C_f、以及非零值V_f。非零值V_f可布置为使得输入数据41的具有相同行索引|R_f^a|的元素彼此邻近。例如,输入数据41的行索引|R_f^a|为0的元素f_{00}、f_{02}和f_{0a}可被称为压缩输入数据$T_f[0]$的第一子张量,输入数据41的行索引|R_f^a|为1的元素f_{14}、f_{16}、f_{1a}和f_{1e}可被称为压缩输入数据$T_f[0]$的第二子张量,输入数据41的行索引|R_f^a|为2的元素f_{20}、f_{22}、f_{28}和f_{2e}可被称为压缩输入数据$T_f[0]$的第三子张量,并且输入数据41的行索引|R_f^a|为3的元素f_{30}、f_{34}、f_{36}和f_{3c}可被称为压缩输入数据$T_f[0]$

的第四子张量。

[0065] 例如,压缩输入数据$T_f[1]$可由输入数据41的多个元素之中的“通过使用图1的等式1而具有非零值并且在多个元素的列索引C_f^o除以2时具有余数值1”的元素组成,2是特定层的步幅。由于余数值是1,因此压缩输入数据$T_f[1]$的索引可以是1。压缩输入数据$T_f[1]$可包括指示输入数据41的行地址的行指针R_f、指示输入数据41的列地址的列索引C_f、以及非零值V_f。非零值V_f可布置为使得输入数据41的具有相同行索引R_f^o的元素彼此邻近。例如,输入数据41的行索引R_f^o为0的元素f_{03}、f_{05}、f_{07}、f_{0b}和f_{0f}可被称为压缩输入数据$T_f[1]$的第一子张量,输入数据41的行索引R_f^o为1的元素f_{11}、f_{19}和f_{1d}可被称为压缩输入数据$T_f[1]$的第二子张量,输入数据41的行索引R_f^o为2的元素f_{21}、f_{25}、f_{27}和f_{2f}可被称为压缩输入数据$T_f[1]$的第三子张量,并且输入数据41的行索引R_f^o为3的元素f_{33}、f_{37}和f_{39}可被称为压缩输入数据$T_f[1]$的第四子张量。

[0066] 图4A是示出根据至少一个实施例的处理电路的基于特定层的步幅通过使用SCSR算法对ANN中的特定层的权重数据进行压缩的操作的示图。特定层的权重数据42可示出以4×4矩阵的形式形成的稀疏矩阵,并且特定层的步幅可以是2。

[0067] 在一些实施例中,进一步参照图1,压缩电路200可基于特定层的步幅通过使用SCSR算法从权重数据42提取非零值来将权重数据42压缩为两个压缩权重数据$T_w[0]$和$T_w[1]$。

[0068] 例如,压缩权重数据$T_w[0]$可由权重数据42的多个元素之中的“通过使用图1的等式1而具有非零值并且在多个元素的列索引C_w^o除以2时具有余数值0”的元素组成,2是特定层的步幅。由于余数值是0,因此压缩权重数据$T_w[0]$的索引可以是0。压缩权重数据$T_w[0]$可包括指示权重数据42的行地址的行指针R_w、指示权重数据42的列地址的列索引C_w、以及非零值V_w。非零值V_w可布置为使得权重数据42的具有相同行索引R_w^o的元素彼此邻近。

[0069] [等式2]

$$r_w = (r_w^o \ mod \ s) \times s + floor(r_w^o / s)$$

在等式2中,r_w是压缩权重数据的行索引值;并且r_w^o是权重数据的行索引值。mod可表示用于计算余数的数学符号,并且$floor$可表示用于向下舍入的数学符号。例如,顺序地布置权重数据42的行索引R_f^o为奇数的元素W_{00}、W_{02}和W_{20},然后布置权重数据42的行索引R_f^o为偶数的元素W_12。

[0070] 例如,压缩权重数据$T_w[1]$可由权重数据42的多个元素之中的“通过使用图1的等式1而具有非零值并且在多个元素的列索引C_w^o除以2时具有余数值1”的元素组成,2是特定层的步幅。由于余数值是1,因此压缩权重数据$T_w[1]$的索引可以是1。压缩权重数据$T_w[1]$可包括指示权重数据42的行地址的行指针R_w、指示权重数据42的列地址的列索引C_w、以及非零值V_w。非零值V_w可布置为使得权重数据42的行索引R_w^o满足以上的等式2。例如,顺序地布置权重数据42的行索引R_f^o为奇数的元素W_{01}、W_{21}和W_{23},然后布置权重数据42的行索引R_f^o为偶数的元素W_{13}、W_{31}和W_{33}。

[0071] 当满足以上的等式2并且执行对图4A的输入数据41和权重数据42的卷积运算时,

可减少或防止边界部分处的无效乘法(例如,f_{00}和W_{20}的乘法)。因此,根据发明构思的处理装置可跳过(例如,不执行)搜索有效输入对的操作。因此,根据发明构思的处理装置可在减少功耗的同时提高处理速度。此外,提高的速度和/或减少的功耗可允许在有限环境(诸如,移动环境和/或需要高速处理的环境)中实现ANN,其中,对比示例将由于相对较大的处理速度和/或功耗而不适合。此外,减少的功耗可减少处理装置中的发热,从而通过缓解和/或延迟由于过热导致的性能劣化来提高处理装置的性能。

[0072] 在一些实施例中,权重数据42可被预压缩为压缩权重数据$T_w[0]$和$T_w[1]$。例如,在ANN的训练中,在生成ANN内的每个层的输出数据之前,可基于特定层的步幅通过使用SCSR算法将权重数据42压缩为压缩权重数据$T_w[0]$和$T_w[1]$。

[0073] 在一些实施例中,输入数据41可在ANN的推断操作中被压缩为压缩输入数据$T_f[0]$和$T_f[1]$。推断操作可表示生成ANN中的每个层的输出数据的操作。

[0074] 图5是根据至少一个实施例的处理电路的框图。

[0075] 参照图5,在一些实施例中,处理电路10a可包括图1的处理电路10的至少一个实施例。处理电路10a可包括计算电路100a、压缩电路200a、输入电路300a、权重缓冲器400a和主控制器500a,其中,压缩电路200a、输入电路300a和权重缓冲器400a可分别与图1的压缩电路200、输入电路300和权重缓冲器400相同(和/或基本相似)。与图1重叠的描述可被省略。

[0076] 计算电路100a可包括以矩阵形式布置的多个处理元件(PE)。在一些实施例中,计算电路100a可包括以矩阵(例如,16×16矩阵)布置的多个PE。在多个PE之中,具有相同的行索引的PE可从输入电路300a接收相同的压缩输入数据,并且具有相同的列索引的PE可从权重缓冲器400a接收相同的压缩权重数据。将相同的数据发送到多个PE的操作可被称为广播。例如,具有行索引0的PE(例如,布置在第一行中的PE)可从输入电路300a接收相同的压缩输入数据,并且具有列索引0的PE(例如,布置在第一列中的PE)可从权重缓冲器400a接收相同的压缩权重数据。由于布置在相同的行或相同的列中的PE接收相同的输入,因此访问存储器(例如,输入电路300a或权重缓冲器400a)的次数可被减少,从而提高PE的处理速度。

[0077] 主控制器500a可从输入电路300a接收压缩输入数据,并且基于压缩输入数据来生成(或计算)压缩权重数据的开始元素索引r_w^{start}和结束元素索引r_w^{end}。为了生成特定层的输出数据,可对特定层的压缩输入数据和压缩权重数据执行卷积运算,其中,卷积运算可表示二维卷积运算。压缩权重数据的开始元素索引r_w^{start}和结束元素索引r_w^{end}可表示压缩权重数据的在二维卷积运算中与特定层的压缩输入数据形成有效输入对的元素的行索引之中的开始索引和结束索引。例如,压缩权重数据的开始元素索引可表示压缩权重数据的有效元素之中的具有最小的行索引的元素的索引,并且压缩权重数据的结束元素索引可表示压缩权重数据的有效元素之中的具有最大的行索引的元素的索引。在一些实施例中,主控制器500a可基于压缩输入数据使用第一算法来计算压缩权重数据的开始元素索引r_w^{start}和结束元素索引r_w^{end}。第一算法可如下表1中所示。

[0078] [表1]

输入： r_f-压缩特征行索引

s-步幅

R_w^o-权重矩阵高度

R_f^o-特征矩阵高度

输出： r_w^{start}、r_w^{end}

1. **如果**（$R_f^o - r_f < R_w^o$）
2. $r_w^{start} = (r_f mods) \times s + floor((R_w^o - R_f^o + r_f)/s)$
3. **否则**
4. $r_w^{start} = (r_f mods) \times s$
5. **结束如果**
6. **如果**（$r_f < R_w^o - 1$）
7. $r_w^{end} = (r_f mods) \times s + floor(r_f/s)$
8. **否则**
9. $r_w^{end} = \left((r_f mods) + 1\right) \times s - 1$
10. **结束如果**

[0079] 第一算法可表示接收输入、执行所有行的计算并且生成输出的操作。

[0080] 主控制器500a可基于压缩输入数据来计算小阵列$Sel[\ ]$。为了生成特定层的输出数据，可对特定层的压缩输入数据和压缩权重数据执行卷积运算，其中，卷积运算可表示二维卷积运算。小阵列$Sel[\ ]$可表示用于选择输入计数器（例如，输入计数器值）cnt_f的初始值的阵列。输入计数器cnt_f的范围可包括在压缩输入数据的开始元素索引c_f^{start}与结束元素索引c_f^{end}之间的范围，其中，压缩输入数据的开始元素索引c_f^{start}和结束元素索引c_f^{end}可分别表示压缩输入数据的在二维卷积运算中与特定层的压缩权重数据形成有效输入对的元素的列索引之中的开始索引和结束索引。例如，压缩输入数据的开始元素索引可表示压缩输入数据的有效元素之中的具有最小的行索引的元素的索引，并且压缩输入数据的结束元素索引可表示压缩输入数据的有效元素之中的具有最大的行索引的元素的索引。

[0081] 在一些实施例中，主控制器500a可基于压缩输入数据使用第二算法来计算小阵列$Sel[\ ]$。第二算法可如下表2中所示。

[0082] [表2]

输入： C_f-特征列索引

C_w^o-权重矩阵长度

输出： $Sel[0: C_w^o - 1]$

1. 整型变量 $cnt = 0$
2. 针对（$i = 0; i < C_w^o - 1; i++$）
3. $Sel[i] = cnt$
4. 如果（$C_f[cnt] == i$）
5. $cnt += 1$
6. 结束如果
7. 结束循环

[0083] 第二算法可表示接收输入、执行所有行的计算并且生成输出的操作。主控制器500a可将计算的小阵列$Sel[\ ]$、以及压缩权重数据的开始元素索引r_w^{start}和结束元素索引r_w^{end}发送给包括在计算电路100a中的多个PE。

[0084] 在一些实施例中,由计算电路100a生成的特定层的输出数据可被用于非线性转换和下采样在其中被执行的激活函数(未示出)和池化层(未示出),并且可被发送到压缩电路200a。激活函数可包括非线性函数(诸如,修正线性单元(ReLU,又称为线性整流函数)、参数化修正线性单元(PReLU)、双曲正切(tanh)函数或sigmoid函数),并且下采样可表示减小数据的大小的操作。例如,特定层的输出数据可被用于激活函数(未示出)和池化层(未示出),被非线性地转换,被减小大小,并且被发送到压缩电路200a。

[0085] 图6是根据至少一个实施例的PE的框图。

[0086] 参照图6,PE 110可包括图5中的多个PE中的每个的配置的至少一个实施例,并且输入电路300b和权重缓冲器400b可分别与图5中的输入电路300a和权重缓冲器400a相同。与图5重叠的描述被省略。PE 110可包括乘法累加(MAC)电路111、地址计算电路112、输出缓冲器113和PE控制器114。

[0087] MAC电路111可通过基于从输入电路300b接收的压缩输入数据和从权重缓冲器400b接收的压缩权重数据执行乘法和累加来生成输出数据。在一些实施例中,MAC电路111可从输入电路300b接收图4A的压缩输入数据$T_f[0]$和$T_f[1]$,并且可从权重缓冲器400b接收图4B的压缩权重数据$T_w[0]$和$T_w[1]$。MAC电路111可通过基于图4A的压缩输入数据$T_f[0]$和$T_f[1]$以及图4B的压缩权重数据$T_w[0]$和$T_w[1]$执行乘法和累加来生成输出数据。

[0088] 地址计算电路112可基于从输入电路300b接收的压缩输入数据的地址信号来计算输出数据的地址。例如,地址计算电路112可计算输出数据的地址以与压缩输入数据的地址信号对应。

[0089] 输出缓冲器113可从地址计算电路112接收输出数据的地址,并且可从MAC电路111接收输出数据。输出缓冲器113可将接收的输出数据的地址映射到输出数据,并且存储输出数据。例如,输出缓冲器113还可存储输出数据的地址。

[0090] PE控制器114可控制输入电路300b将压缩输入数据的有效元素发送到MAC电路

111,并且可控制权重缓冲器400b将压缩权重数据的有效元素发送到MAC电路111。

[0091] 在一些实施例中,PE控制器114可从图5的主控制器500a接收小阵列$Sel[\ ]$以及压缩权重数据的开始元素索引r_w^{start}和结束元素索引r_w^{end},并且可通过基于小阵列$Sel[\ ]$以及压缩权重数据的开始元素索引r_w^{start}和结束元素索引r_w^{end}使用卷积算法来计算输入计数器cnt_f和权重计数器(例如,权重计数器值)cnt_w。卷积算法可如下表3中所示。

[0092] [表3]

输入: r_w^{start}、r_w^{end}
$Sel[0: C_w^o - 1]$
C_f^o、C_w^o
T_w
输出: cnt_w、cnt_f

1. $cnt_w = find_index(T_w, r_w^{start})$
2. **当** $cnt_w \le find_index(T_w, r_w^{end})$ **时**
3. $r_w = find_column(T_w, cnt_w)$
4. $cnt_f = Sel[r_w]$
5. **当** $cnt_f \le C_f^o - C_w^o + r_w$ **时**
6. $cnt_f += 1$
7. **结束循环**
8. $cnt_w += 1$
9. **结束循环**

[0093] 卷积算法可表示接收输入、执行所有行的计算并且生成输出的操作。

[0094] 在一些实施例中,PE控制器114可被配置为控制输入电路300b和权重缓冲器400b,以基于输入计数器cnt_f和权重计数器cnt_w将作为有效输入对的压缩输入数据的元素和压缩权重数据的元素发送到MAC电路111。以下可参照图8描述PE控制器114基于输入计数器cnt_f和权重计数器cnt_w来控制输入电路300b和权重缓冲器400b的特定实施例。

[0095] 图7是根据至少一个实施例的处理电路的卷积运算的流程图。图8是示出根据至少一个实施例的处理电路的卷积运算的示图。参照图7,处理电路的卷积运算70可包括多个操作S710至S730。进一步参照图8,图8可表示根据时钟的图4A的压缩输入数据$T_f[0]$的元素和图4B的压缩权重数据$T_w[0]$的元素的卷积运算。

[0096] 进一步参照图5和图6,在操作S710中,处理电路10a可计算输入计数器cnt_f和权重计数器cnt_w。在一些实施例中,PE控制器114可从主控制器500a接收小阵列$Sel[\ ]$以及压缩权重数据的开始元素索引r_w^{start}和结束元素索引r_w^{end},并且可通过使用表3来计算

输入计数器cnt_f和权重计数器cnt_w,表3是基于小阵列$Sel[\quad]$以及压缩权重数据的开始元素索引r_w^{start}和结束元素索引r_w^{end}的卷积算法。

[0097] 在操作S720中,MAC电路111可基于权重计数器cnt_w接收压缩权重数据的元素之中的有效元素,并且基于输入计数器cnt_f接收压缩输入数据的元素之中的有效元素。

[0098] 在一些实施例中,PE控制器114可控制权重缓冲器400b,使得MAC电路111基于权重计数器cnt_w接收压缩权重数据的元素之中的有效元素,并且PE控制器114可控制输入电路300b,使得MAC电路111基于输入计数器cnt_f接收压缩输入数据的元素之中的有效元素。

[0099] 例如,MAC电路111可从权重缓冲器400b接收压缩权重数据$T_w[0]$的元素W_{00},并且PE控制器114可控制输入电路300b,使得MAC电路111基于输入计数器cnt_f接收压缩输入数据$T_f[0]$的与元素W_{00}形成有效输入对的元素f_{00}、f_{02}和f_{0a}。当不存在压缩输入数据的与元素W_{00}形成有效输入对的元素时,PE控制器114可控制权重缓冲器400b,使得MAC电路111基于权重计数器cnt_w在第三时钟(时钟2)接收压缩权重数据$T_w[0]$的有效元素之中的元素W_{00}的后续元素W_{02}。PE控制器114可控制输入电路300b,使得MAC电路111基于输入计数器cnt_f接收压缩输入数据$T_f[0]$的与元素W_{02}形成有效输入对的元素f_{02}和f_{0a}。当不存在压缩输入数据的与元素W_{02}形成有效输入对的元素时,PE控制器114可控制权重缓冲器400b,使得MAC电路111基于权重计数器cnt_w在第五时钟(时钟4)接收压缩权重数据$T_w[0]$的有效元素之中的元素W_{02}的后续元素W_{12}。换言之,PE控制器114可顺序地控制权重缓冲器400b和输入电路300b,使得MAC电路111接收有效输入对。

[0100] 在操作S730中,MAC电路111可通过基于有效元素执行卷积运算来生成输出数据。例如,在图8中,可通过在从第一时钟(时钟0)到第21时钟(时钟20)的每个时钟(例如,每个时钟周期)对有效输入对执行乘法并对乘法结果进行累加来生成输出数据。

[0101] 图9示出将根据至少一个实施例的处理电路与根据对比实施例的处理电路在处理电路的每个组件的功耗和功耗百分比方面进行比较的图。

[0102] 参照图9,图91将对比实施例的处理电路通过使用压缩算法(A)(例如,CSR)对特定层的输入数据或权重数据进行压缩时的功耗与根据至少一个实施例的处理电路基于特定层的步幅通过使用SCSR算法(B)对特定层的输入数据或权重数据进行压缩时的功耗进行比较。图91的纵轴可表示功耗,并且图91的横轴可表示压缩方法。可看出,使用压缩算法(A)时的功耗比使用SCSR算法(B)时大4倍。换言之,当使用SCSR算法(B)时,处理电路的功耗可被减少。

[0103] 图92示出根据每种压缩方法的处理电路的每个组件的功耗百分比。由于处理电路在使用压缩算法(A)时需要搜索有效输入对的操作,因此功耗的显著部分可由单独电路(例如,前缀求和或优先级编码器)占用以搜索有效输入对。相反,由于处理电路在使用SCSR算法(B)时不需要搜索有效输入对的操作,因此功耗的显著部分可被分配给缓冲器(例如,输出缓冲器、输入缓冲器和权重缓冲器)。相应地,当使用SCSR算法(B)时,搜索有效输入对的操作被省略,从而减少功耗并提高处理速度。

[0104] 图10是根据至少一个实施例的计算系统的框图。

[0105] 在一些实施例中,图1的处理电路10可被实现为图10的计算系统2000。如图10中所示,计算系统2000可包括系统存储器2100、处理器2300、存储设备2500、输入/输出(IO)装置

2700和通信连接2900。包括在计算系统2000中的组件可例如通过总线通信地彼此连接。

[0106] 系统存储器2100可包括程序2120。根据一些实施例,程序2120可使得处理器2300执行ANN的量化。例如,程序2120可包括可由处理器2300执行的多个指令。当包括在程序2120中的多个指令由处理器2300执行时,每个层的输出数据可基于ANN中的每个层的步幅通过使用SCSR算法被压缩为后续层的输入数据。作为一个非限制性示例,系统存储器2100可包括易失性存储器(诸如,SRAM和DRAM)和非易失性存储器(诸如,闪存)。

[0107] 处理器2300可包括能够执行任何指令集的至少一个核(例如,英特尔架构-32(IA-32)、64位扩展IA-32、x86-64、PowerPC、Sparc、无互锁流水线级的微处理器(MIPS)、高级精简指令集计算机(RISC)机器(ARM)、IA-64等)。处理器2300可执行存储在系统存储器2100中的指令,并且可基于ANN中的每个层的步幅通过使用SCSR算法将每个层的输出数据压缩为后续层的输入数据。

[0108] 即使供应给计算系统2000的电力被切断,存储设备2500也可不丢失存储的数据。例如,存储设备2500可包括非易失性存储器(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)、电阻式随机存取存储器(RRAM)、纳米浮栅存储器(NFGM)、聚合物随机存取存储器(PoRAM)、磁式随机存取存储器(MRAM)或铁电随机存取存储器(FRAM)),并且还可包括存储介质(诸如,磁带、光盘或磁盘)。在一些实施例中,存储设备2500可以是可从计算系统2000移除的。

[0109] 在一些实施例中,存储设备2500可存储用于基于根据至少一个实施例的ANN中的每个层的步幅通过使用SCSR算法将每个层的输出数据压缩为后续层的输入数据的程序2120,其中,程序2120或程序2120的至少一部分可在程序2120由处理器2300执行之前从存储设备2500被加载到系统存储器2100中。在一些实施例中,存储设备2500可存储以程序语言编写的文件,并且由编译器等从该文件生成的程序2120或程序2120的至少一部分可被加载到系统存储器2100中。

[0110] 输入/输出装置2700可被配置为包括输入装置(诸如,键盘和定点装置)和输出装置(诸如,显示装置和打印机)。例如,用户可通过输入/输出装置2700触发处理器2300对程序2120的执行。

[0111] 通信连接2900可被配置为提供对计算系统2000外部的网络的访问。例如,网络可包括多个计算系统和通信链路,其中,通信链路可包括有线链路、光学链路、无线链路或任何其他类型的链路。

[0112] 图11是根据至少一个实施例的便携式计算装置的框图。

[0113] 在一些实施例中,用于基于ANN中的每个层的步幅通过使用SCSR算法将每个层的输出数据压缩为后续层的输入数据的电路可实现在便携式计算装置(例如,移动计算装置)3000中。作为一个非限制性示例,便携式计算装置3000可包括由电池或自生电力供电的任何便携式电子装置(诸如,移动电话、平板PC、可穿戴装置、物联网装置等)。

[0114] 如图11中所示,便携式计算装置3000可包括存储器子系统3100、输入/输出装置3300、处理单元3500和网络接口3700,其中,存储器子系统3100、输入/输出装置3300、处理单元3500和网络接口3700可通过总线3900彼此通信。在一些实施例中,存储器子系统3100、输入/输出装置3300、处理单元3500和网络接口3700中的至少两个可作为片上系统(SoC)被包括在一个封装件中。

[0115] 存储器子系统3100可包括RAM 3120和存储设备3140。RAM 3120和/或存储设备3140可存储由处理单元3500执行的指令和由处理单元3500处理的数据。例如,RAM 3120和/或存储设备3140可存储变量(诸如,ANN的信号、权重和偏置),并且可存储ANN的人工神经元(或计算节点)的参数。在一些实施例中,存储设备3140可包括非易失性存储器。

[0116] 处理单元3500可包括中央处理器(CPU)3520、图形处理器(GPU)3540、数字信号处理器(DSP)3560和神经处理器(NPU)3580。在一些实施例中,与图10中所示的内容不同,处理单元3500可包括CPU 3520、GPU 3540、DSP 3560和NPU 3580中的至少一些。

[0117] CPU 3520可执行便携式计算装置的整体操作(例如,响应于通过输入/输出装置3300接收的外部输入而直接执行特定任务,或者指示处理单元3500的其他组件执行任务)。GPU 3540可生成用于通过包括在输入/输出装置3300中的显示装置输出的图像的数据,并且可对从包括在输入/输出装置3300中的相机接收的数据进行编码。DSP 3560可通过对数字信号(例如,从网络接口3700提供的数字信号)进行处理来生成有用的数据。

[0118] 作为用于ANN的专用硬件的NPU 3580可包括与构成ANN的人工神经元中的至少部分对应的多个计算节点,其中,多个计算节点中的至少部分可并行地对信号进行处理。处理单元3500可包括根据上述实施例的用于基于ANN中的每个层的步幅使用SCSR算法将每个层的输出数据压缩为后续层的输入数据的电路。

[0119] 输入/输出装置3300可包括输入装置(诸如,触摸输入装置、声音输入装置和相机)和输出装置(诸如,显示装置和声音输出装置)。网络接口3700可向便携式计算装置3000提供对移动通信网络(诸如,长期演进(LTE)或5G)的访问或对本地网络(诸如,Wi-Fi)的访问。

[0120] 虽然已经参照发明构思的实施例具体示出和描述了发明构思,但是将理解,在不脱离所附权利要求的精神和范围的情况下,可在其中进行形式和细节上的各种改变。

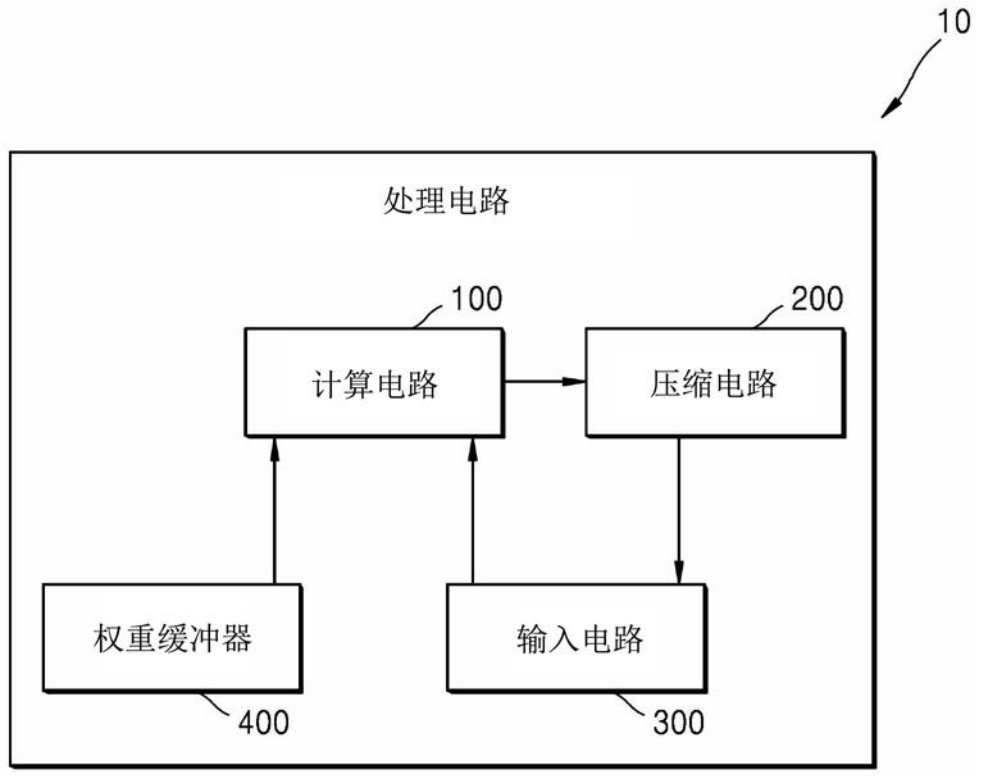

10
处理电路
100
计算电路
200
压缩电路
权重缓冲器
400
输入电路
300

图1

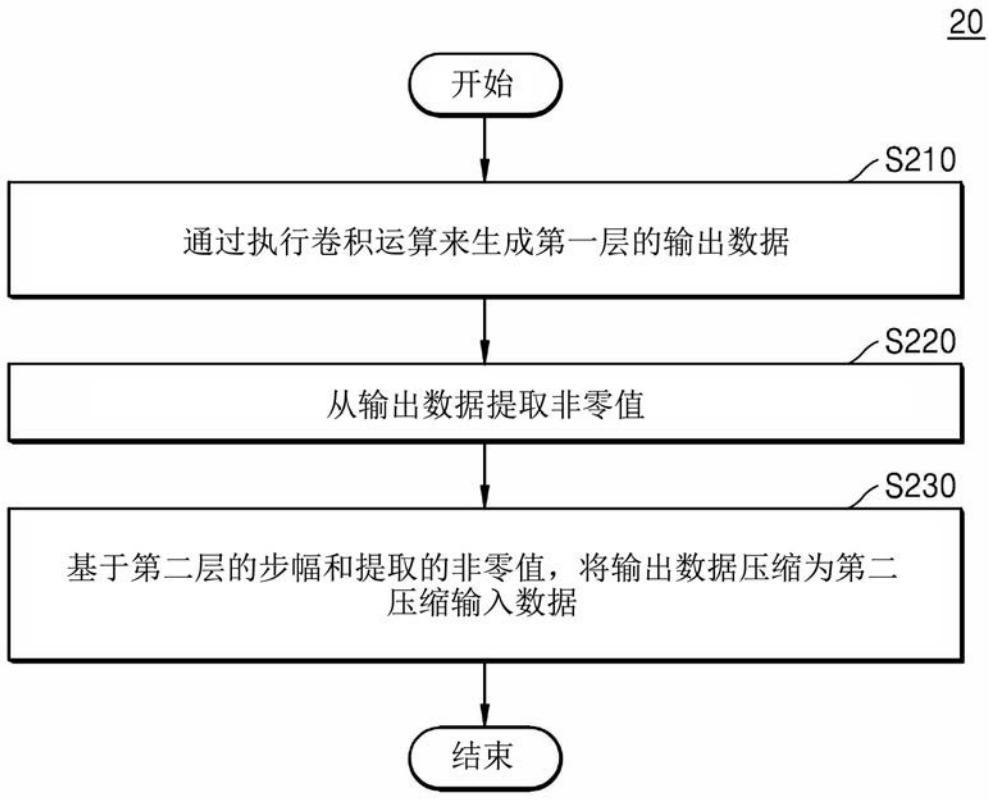

20
开始
S210
通过执行卷积运算来生成第一层的输出数据
S220
从输出数据提取非零值
S230
基于第二层的步幅和提取的非零值，将输出数据压缩为第二压缩输入数据
结束

图2

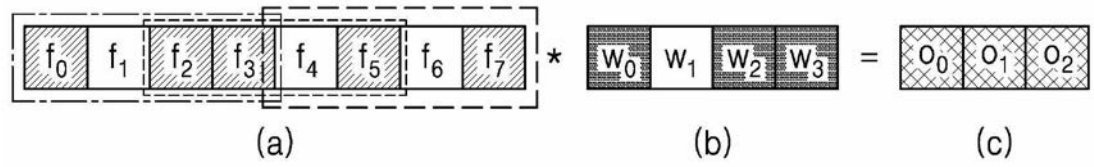


图3A

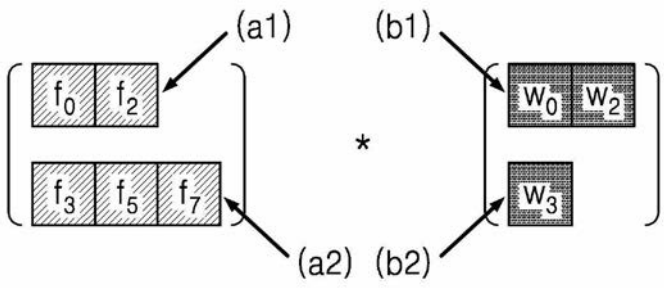


图3B

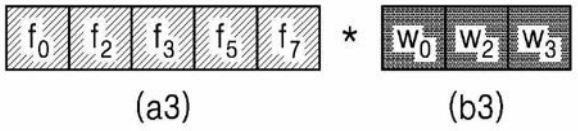


图3C

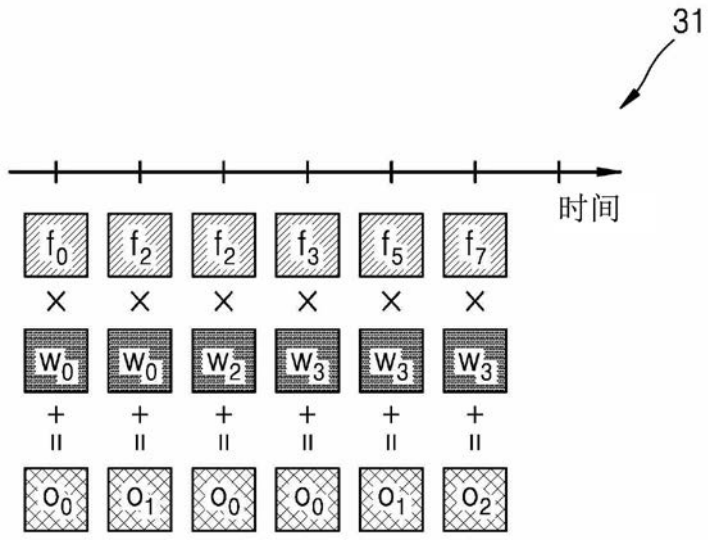


图3D

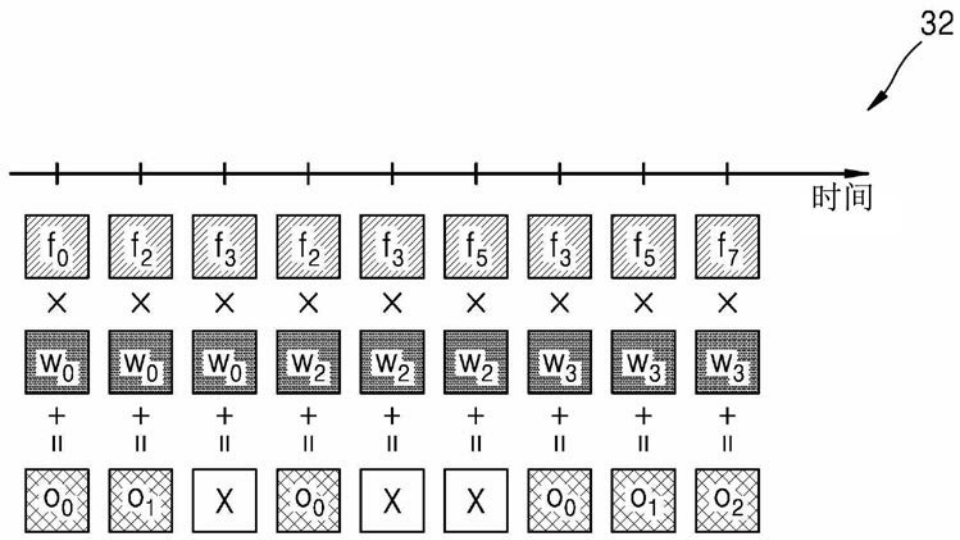

32
时间
f0 f2 f3 f2 f3 f5 f3 f5 f7
w0 w0 w0 w2 w2 w2 w3 w3 w3
o0 o1 X o0 X X o0 o1 o2

图3E

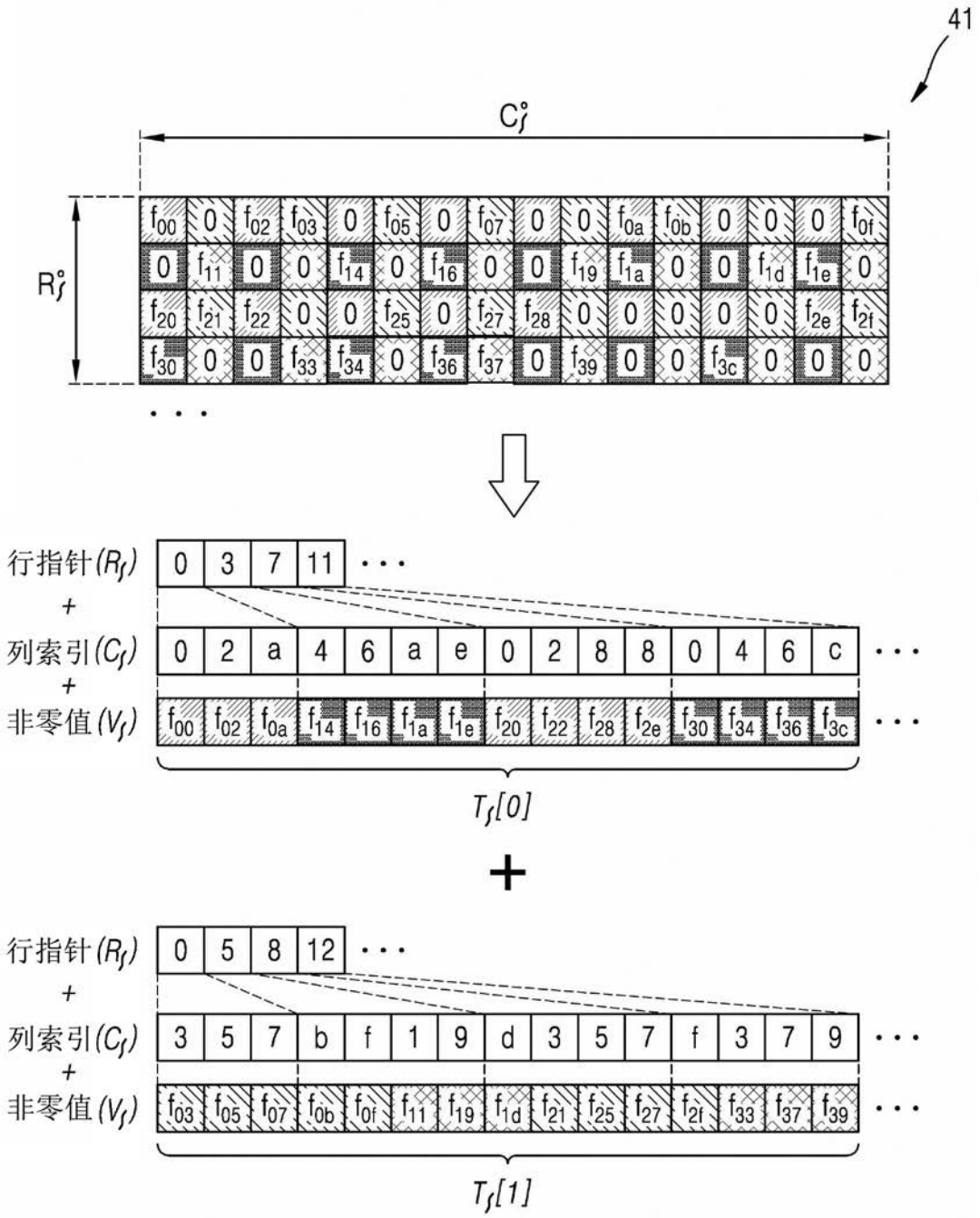

41
C_f°
R_f°
f_{00} 0 f_{02} f_{03} 0 f_{05} 0 f_{07} 0 0 f_{0a} f_{0b} 0 0 0 f_{0f}
0 f_{11} 0 0 f_{14} 0 f_{16} 0 0 f_{19} f_{1a} 0 0 f_{1d} f_{1e} 0
f_{20} f_{21} f_{22} 0 0 f_{25} 0 f_{27} f_{28} 0 0 0 0 0 f_{2e} f_{2f}
f_{30} 0 0 f_{33} f_{34} 0 f_{36} f_{37} 0 f_{39} 0 0 f_{3c} 0 0 0
行指针(R_f) 0 3 7 11 ···
+
列索引(C_f) 0 2 a 4 6 a e 0 2 8 8 0 4 6 c ···
+
非零值(V_f) f_{00} f_{02} f_{0a} f_{14} f_{16} f_{1a} f_{1e} f_{20} f_{22} f_{28} f_{2e} f_{30} f_{34} f_{36} f_{3c} ···
$T_f[0]$
+
行指针(R_f) 0 5 8 12 ···
+
列索引(C_f) 3 5 7 b f 1 9 d 3 5 7 f 3 7 9 ···
+
非零值(V_f) f_{03} f_{05} f_{07} f_{0b} f_{0f} f_{11} f_{19} f_{1d} f_{21} f_{25} f_{27} f_{2f} f_{33} f_{37} f_{39} ···
$T_f[1]$

图4A

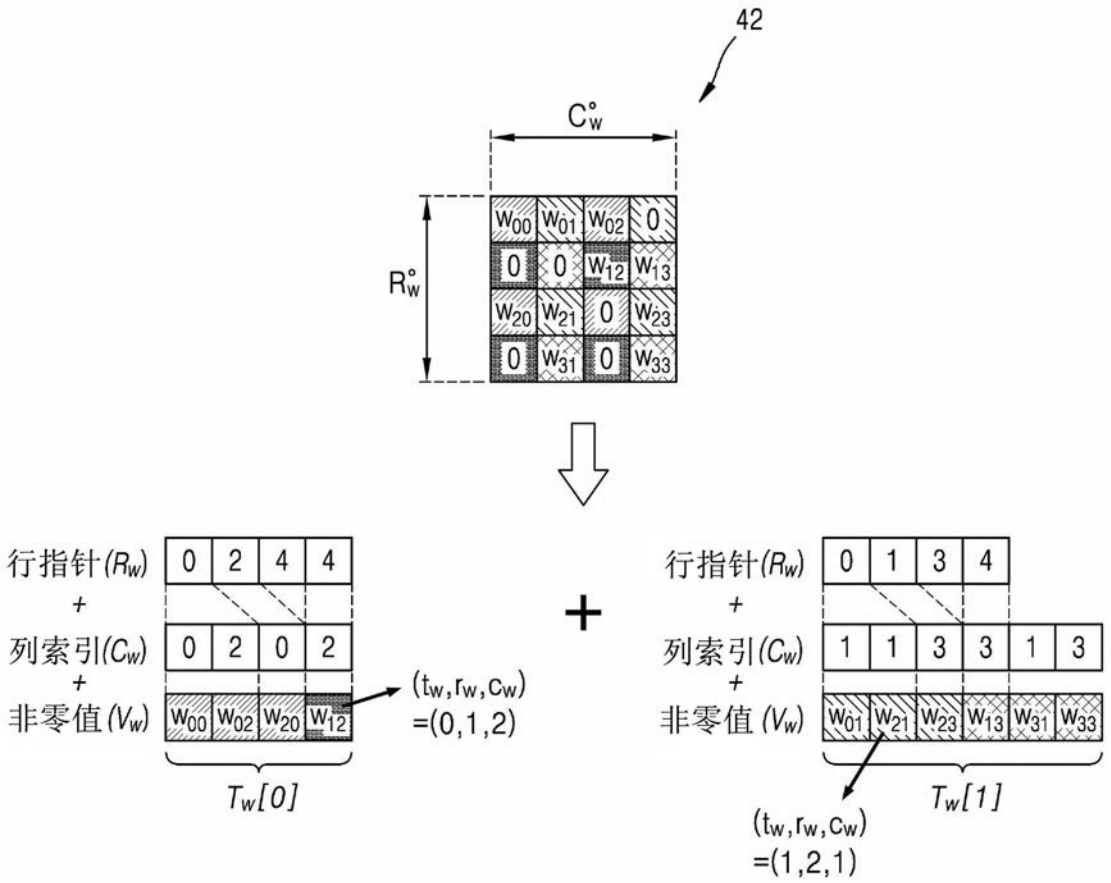

42
C°_{w}
R°_{w}
w_{00} w_{01} w_{02} 0
0 0 w_{12} w_{13}
w_{20} w_{21} 0 w_{23}
0 w_{31} 0 w_{33}
行指针(R_w) 0 2 4 4
+
列索引(C_w) 0 2 0 2
+
非零值(V_w) w_{00} w_{02} w_{20} w_{12}
(t_w, r_w, c_w) =(0,1,2)
$T_w[0]$
+
行指针(R_w) 0 1 3 4
+
列索引(C_w) 1 1 3 3 1 3
+
非零值(V_w) w_{01} w_{21} w_{23} w_{13} w_{31} w_{33}
$T_w[1]$
(t_w, r_w, c_w) =(1,2,1)

图4B

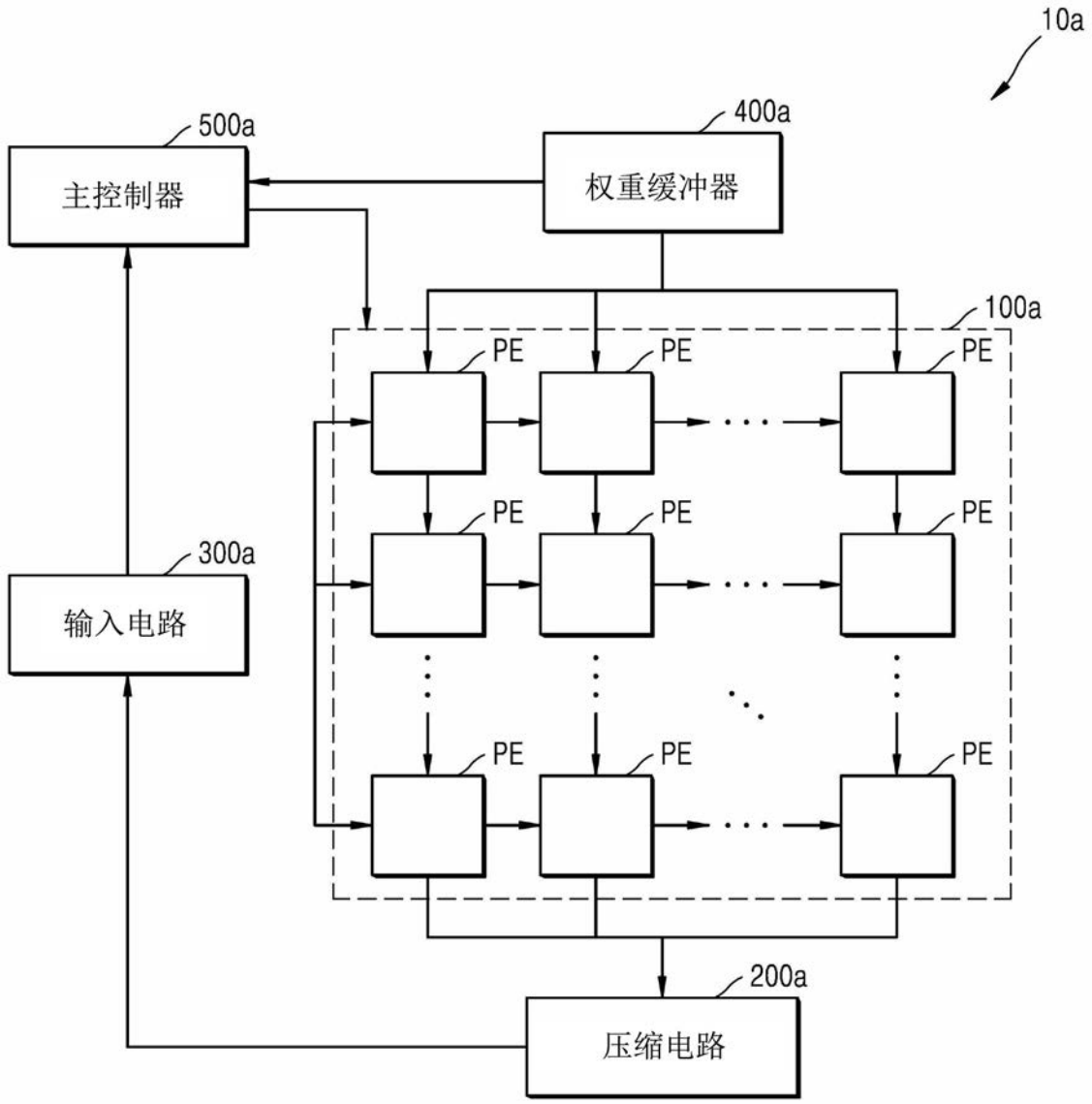


图5

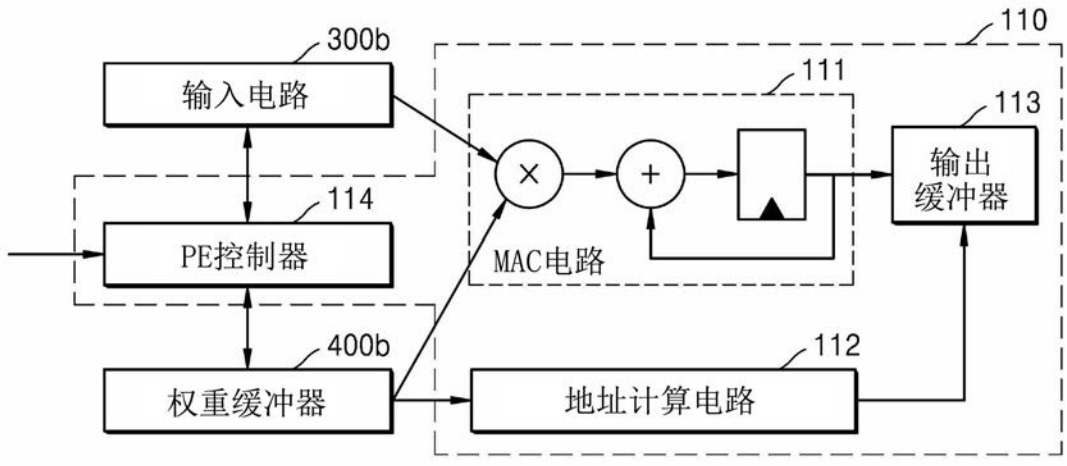

110
300b
输入电路
111
113
输出
缓冲器
114
PE控制器
MAC电路
400b
112
权重缓冲器
地址计算电路

图6

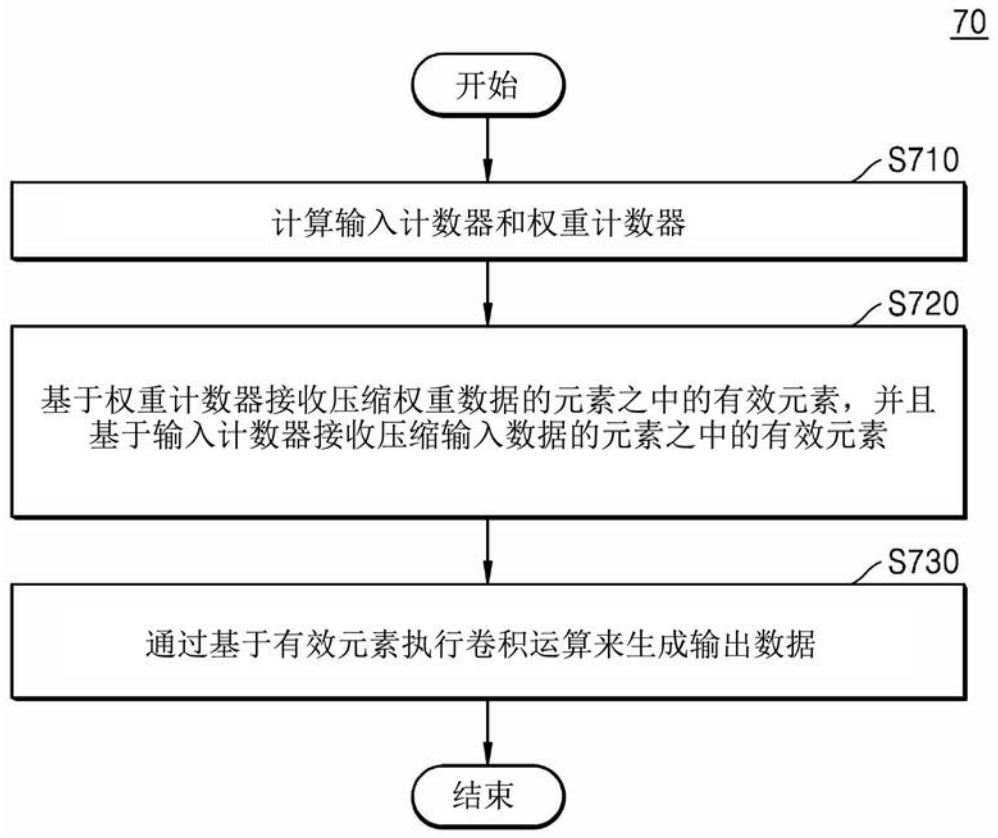

70
开始
S710
计算输入计数器和权重计数器
S720
基于权重计数器接收压缩权重数据的元素之中的有效元素，并且
基于输入计数器接收压缩输入数据的元素之中的有效元素
S730
通过基于有效元素执行卷积运算来生成输出数据
结束

图7

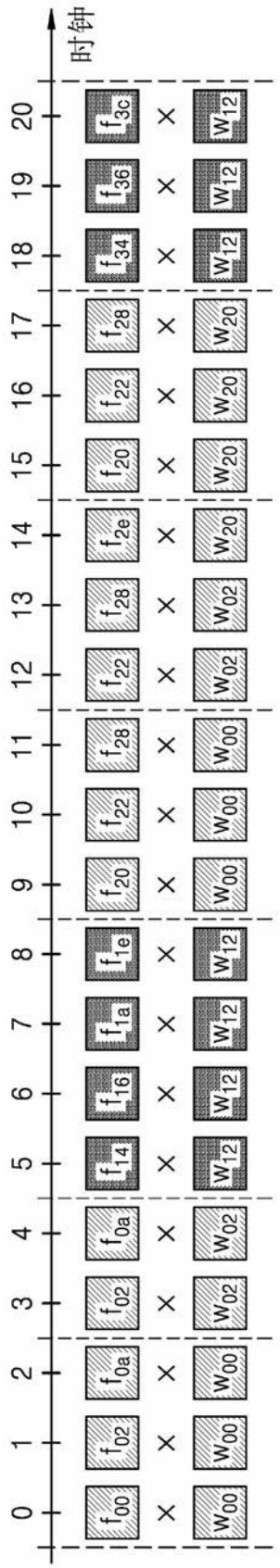

时钟
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
f00
f02
f0a
f02
f0a
f14
f16
f1a
f1e
f20
f22
f28
f22
f28
f2e
f20
f22
f28
f34
f36
f3c
w00
w00
w00
w02
w02
w12
w12
w12
w12
w00
w00
w00
w02
w02
w20
w20
w20
w12
w12
w12
×

图8

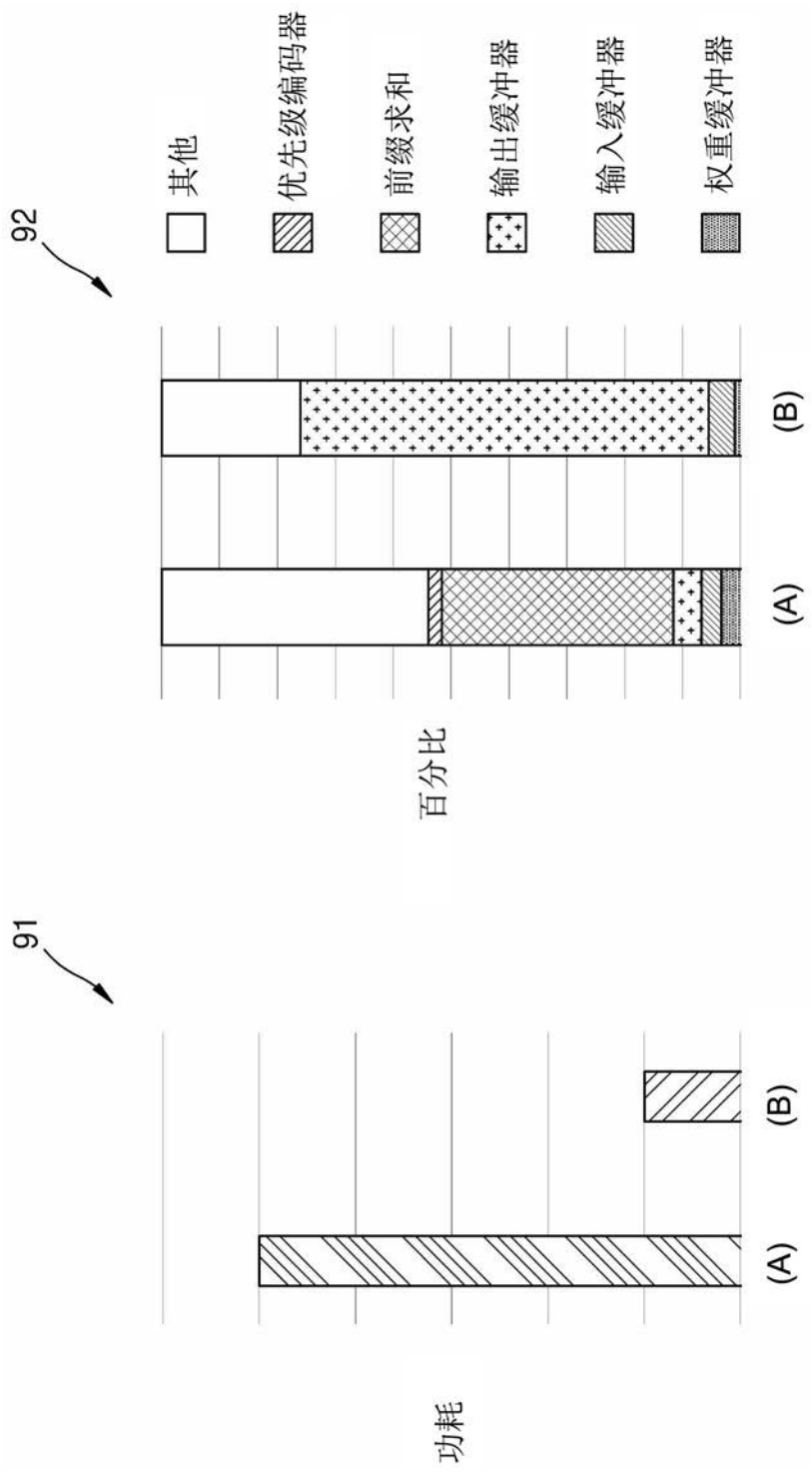


图9

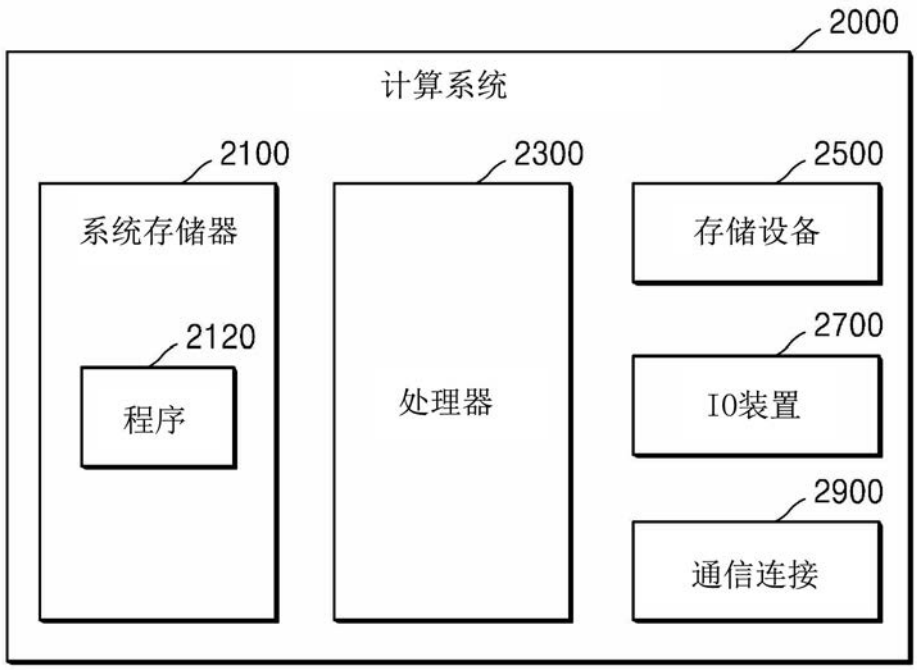

2000
计算系统
2100
系统存储器
2120
程序
2300
处理器
2500
存储设备
2700
IO装置
2900
通信连接

图10

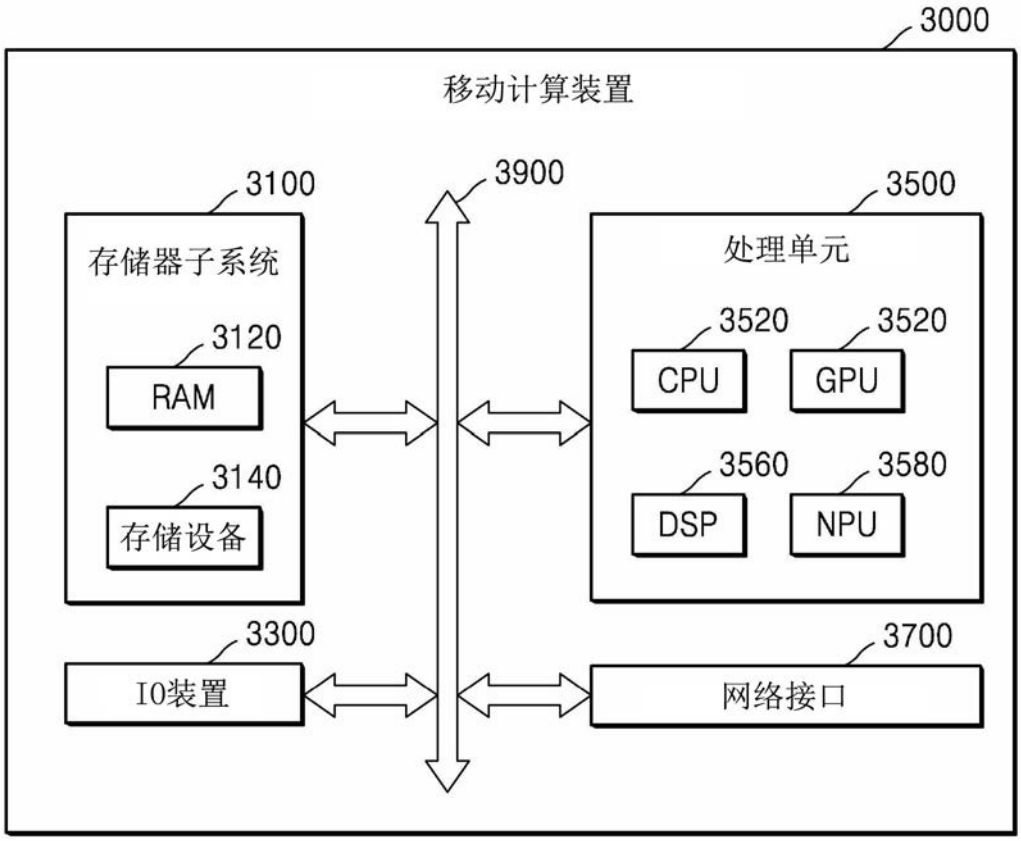

3000
移动计算装置
3100
存储器子系统
3120
RAM
3140
存储设备
3900
3500
处理单元
3520
CPU
3520
GPU
3560
DSP
3580
NPU
3300
IO装置
3700
网络接口

图11