(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111291874 B
(45)授权公告日 2023.12.01

(21)申请号 201910953878.X

(22)申请日 2019.10.09

(65)同一申请的已公布的文献号
申请公布号 CN 111291874 A

(43)申请公布日 2020.06.16

(30)优先权数据
62/776,426 2018.12.06 US

(73)专利权人 神盾股份有限公司
地址 中国台湾台北市内湖区瑞光路360号2楼

(72)发明人 黄朝宗

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

(51)Int.Cl.
G06N 3/063(2023.01)
G06N 3/0464(2023.01)

(56)对比文件
CN 108763191 A,2018.11.06
US 2018189981 A1,2018.07.05
WO 2017177446 A1,2017.10.19
CN 105681628 A,2016.06.15
TW 201706871 A,2017.02.16
TW 201706872 A,2017.02.16
JP H03251947 A,1991.11.11

审查员 吴秀萍

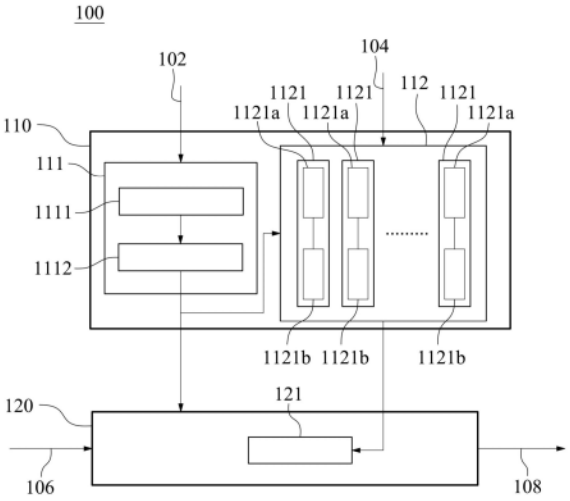
权利要求书5页 说明书16页 附图10页

(54)发明名称

卷积神经网络处理器及其数据处理方法

(57)摘要

一种卷积神经网络处理器包含信息解码单元以及卷积判断单元。信息解码单元接收输入程序及多个输入权重参数并包含解码模块及平行处理模块。解码模块接收输入程序并根据输入程序输出运行指令。平行处理模块与解码模块电性连接并接收输入权重参数。平行处理模块包含多个平行处理子模块。平行处理子模块根据运行指令及输入权重参数产生多个输出权重参数。卷积判断单元与信息解码单元电性连接并包含运算模块。运算模块与平行处理模块电性连接并依据输入数据与输出权重参数运算而产生输出数据。因此,卷积神经网络处理器可执行高度平行运算。



CN 111291874 B

1.一种卷积神经网络处理器，用以运算一输入数据，其特征在于该卷积神经网络处理器包含：

一信息解码单元，用以接收一输入程序及多个输入权重参数，且包含：

一解码模块，接收该输入程序，并根据该输入程序输出一运行指令；及

一平行处理模块，与该解码模块电性连接，并接收所述输入权重参数，且该平行处理模块包含多个平行处理子模块，所述平行处理子模块根据该运行指令及所述输入权重参数产生多个输出权重参数，其中，

所述输入权重参数包含多个第一输入权重参数；

所述输出权重参数包含多个第一输出权重参数；及

所述平行处理子模块包含：

多个平行子存储器，平行地存储所述输入权重参数，所述平行子存储器包含：

多个第一平行子存储器，分别且平行地接收并存储一该第一输入权重参数；及

多个平行子处理器，分别与该解码模块及一该平行子存储器电性连接，所述平行子处理器包含：

多个第一平行子处理器，分别与一该第一平行子存储器电性连接，根据该运行指令接收一该第一输入权重参数，以输出一该第一输出权重参数，

其中各该第一输出权重参数包含多个3×3权重参数；以及

一卷积判断单元，与该信息解码单元电性连接，且包含：

一运算模块，与该平行处理模块电性连接，该运算模块依据该输入数据与所述输出权重参数运算而产生一输出数据，该运算模块包含：

一3×3运算子模块，与所述第一平行子处理器电性连接，并根据所述第一输出权重参数与该输入数据进行运算，以产生一3×3后处理运算数据，该3×3运算子模块包含：

多个3×3卷积分配器组，各该3×3卷积分配器组与一该第一平行子处理器电性连接，所述3×3卷积分配器组用以接收及分配所述第一输出权重参数的所述3×3权重参数；

多个3×3本地卷积运算单元，分别与一该3×3卷积分配器组电性连接，各该3×3本地卷积运算单元包含：

一3×3本地暂存器组，该3×3本地暂存器组与一该3×3卷积分配器组电性连接，所述3×3本地卷积运算单元的所述3×3本地暂存器组接收并存储所述第一输出权重参数的所述3×3权重参数，并根据所述第一输出权重参数的所述3×3权重参数，以输出多个3×3运算参数；及

一3×3本地滤波运算单元，与该3×3本地暂存器组电性连接，所述3×3本地卷积运算单元的所述3×3本地滤波运算单元根据所述3×3运算参数与该输入数据进行运算以产生多个3×3运算数据；及

多个3×3后处理运算单元，与所述3×3本地卷积运算单元电性连接，并依据所述3×3运算数据进行一3×3后处理运算，以产生该3×3后处理运算数据；

其中，该输出数据为该3×3后处理运算数据。

2.如权利要求1所述的卷积神经网络处理器，其中该解码模块包含：

一程序存储器，存储该输入程序；及

一指令解码器，与该程序存储器电性连接，该指令解码器将该输入程序解码以输出一

运行指令。

3.如权利要求1所述的卷积神经网络处理器,其中当所述输入权重参数为多个非压缩输入权重参数,所述平行处理子模块包含:

所述多个平行子存储器,平行地存储所述非压缩输入权重参数;及

所述多个平行子处理器,分别与该解码模块及一该平行子存储器电性连接,所述平行子处理器根据该运行指令平行地接收所述非压缩输入权重参数,产生所述输出权重参数。

4.如权利要求1所述的卷积神经网络处理器,其中当所述输入权重参数为多个压缩输入权重参数,所述平行处理子模块包含:

所述多个平行子存储器,平行地存储所述压缩输入权重参数;及

所述多个平行子处理器,分别与该解码模块及一该平行子存储器电性连接,所述平行子处理器根据该运行指令平行地接收并解压缩所述压缩输入权重参数,产生所述输出权重参数。

5.如权利要求1所述的卷积神经网络处理器,其中各该3×3本地暂存器组包含:

二子3×3本地暂存器组,交替地存储一该3×3权重参数或输出该3×3运算参数给该3×3本地滤波运算单元。

6.如权利要求1所述的卷积神经网络处理器,其中,

所述输入权重参数还包含一偏压输入权重参数;

所述输出权重参数还包含一偏压输出权重参数;及

所述平行子存储器还包含:

一偏压平行子存储器,平行地存储该偏压输入权重参数;

所述平行子处理器还包含:

一偏压平行子处理器,与该偏压平行子存储器电性连接,根据该运行指令接收该偏压输入权重参数,以输出该偏压输出权重参数。

7.如权利要求6所述的卷积神经网络处理器,其中,

该偏压输出权重参数包含多个偏压权重参数;及

该运算模块还包含:

一偏压分配器,与该偏压平行子处理器、该3×3运算子模块电性连接,该偏压分配器根据该偏压输出权重参数以产生多个3×3偏压权重参数,并将所述3×3偏压权重参数输出至所述3×3后处理运算单元。

8.如权利要求1所述的卷积神经网络处理器,其中,

所述输入权重参数还包含至少一第二输入权重参数;

所述输出权重参数还包含至少一第二输出权重参数;及

所述平行子存储器还包含:

至少一第二平行子存储器,分别且平行地接收并存储该至少一第二输入权重参数;

所述平行子处理器还包含:

至少一第二平行子处理器,分别与该至少一第二平行子存储器电性连接,根据该运行指令接收该至少一第二输入权重参数,以输出该至少一第二输出权重参数。

9.如权利要求8所述的卷积神经网络处理器,其中该运算模块包含:

一3×3运算子模块,与所述第一平行子处理器电性连接,并根据所述第一输出权重参

数与该输入数据进行运算，以产生一3×3后处理运算数据；及

一1×1运算子模块，与该至少一第二平行子处理器及该3×3运算子模块电性连接，并根据该至少一第二输出权重参数与该3×3后处理运算数据进行运算，以产生一1×1后处理运算数据；

其中，该输出数据为该1×1后处理运算数据。

10.如权利要求9所述的卷积神经网络处理器，其中该至少一第二输出权重参数包含多个1×1权重参数；

该1×1运算子模块包含：

至少一1×1卷积分配器组，与该至少一第二平行子处理器电性连接，用以接收及分配该至少一第二输出权重参数的所述1×1权重参数；

多个1×1本地卷积运算单元，与该至少一1×1卷积分配器电性连接，各该1×1本地卷积运算单元包含：

一1×1本地暂存器组，该1×1本地暂存器组与该至少一1×1卷积分配器组电性连接，所述1×1本地卷积运算单元的该1×1本地暂存器组接收并存储该至少一第二输出权重参数的所述1×1权重参数，并根据该至少一第二输出权重参数的所述1×1权重参数，以输出多个1×1运算参数；及

一1×1本地滤波运算单元，与该1×1本地暂存器组电性连接，所述1×1本地卷积运算单元的该1×1本地滤波运算单元根据所述1×1运算参数与该3×3后处理运算数据进行运算以产生多个1×1运算数据；及

多个1×1后处理运算单元，与所述1×1本地卷积运算单元电性连接，并依据所述1×1运算数据进行一1×1后处理运算，以产生该1×1后处理运算数据。

11.如权利要求10所述的卷积神经网络处理器，其中各该1×1本地暂存器组包含：

二子1×1本地暂存器组，交替地存储一该1×1权重参数或输出该1×1运算参数给该1×1本地滤波运算单元。

12.如权利要求10所述的卷积神经网络处理器，其中，

所述输入权重参数还包含一偏压输入权重参数；

所述输出权重参数还包含一偏压输出权重参数；及

所述平行子存储器还包含：

一偏压平行子存储器，平行地存储该偏压输入权重参数；及

所述平行子处理器还包含：

一偏压平行子处理器，与该偏压平行子存储器电性连接，根据该运行指令接收该偏压输入权重参数，以输出该偏压输出权重参数。

13.如权利要求12所述的卷积神经网络处理器，其中该偏压输出权重参数包含多个偏压权重参数；

该运算模块还包含：

一偏压分配器，与该偏压平行子处理器、该3×3运算子模块及该1×1运算子模块电性连接，该偏压分配器根据该偏压输出权重参数以产生多个3×3偏压权重参数及多个1×1偏压权重参数；

其中，该偏压分配器将所述3×3偏压权重参数输出至所述3×3后处理运算单元；

其中，该偏压分配器将所述1×1偏压权重参数输出至所述1×1后处理运算单元。

14.一种卷积神经网络处理器的数据处理方法，其特征在于包含：

一接收步骤，驱动一信息解码单元接收一输入程序及多个输入权重参数，其中该信息解码单元包含一解码模块及一平行处理模块；

一指令解码步骤，驱动该解码模块接收该输入程序，并根据该输入程序，以产生一运行指令；

一平行处理步骤，驱动该平行处理模块接收所述输入权重参数，并根据该运行指令以平行地处理所述输入权重参数，以产生多个输出权重参数，其中，所述输出权重参数包含多个第一输出权重参数，各该第一输出权重参数包含多个3×3权重参数；以及

一运算步骤，驱动一运算模块接收一输入数据及所述输出权重参数，并根据该运行指令以将该输入数据与所述输出权重参数进行运算，以产生一输出数据，其中，该运算模块包含一3×3运算子模块，该3×3运算子模块包含多个3×3卷积分配器组、多个3×3本地卷积运算单元及多个3×3后处理运算单元，该运算步骤包含：

一第一运算子步骤，驱动该3×3运算子模块接收该输入数据及所述第一输出权重参数，以产生一3×3后处理运算数据，该第一运算子步骤包含：

一3×3参数分配程序，驱动所述3×3卷积分配器组接收所述第一输出权重参数的所述3×3权重参数，并将所述第一输出权重参数的所述3×3权重参数分配至所述3×3本地卷积运算单元，其中各该3×3本地卷积运算单元包含一3×3本地暂存器组及一3×3本地滤波运算单元；

一3×3运算参数产生程序，驱动所述3×3本地卷积运算单元的所述3×3本地暂存器组接收所述第一输出权重参数的所述3×3权重参数，并根据所述第一输出权重参数的所述3×3权重参数产生多个3×3运算参数；

一3×3卷积运算程序，驱动所述3×3本地卷积运算单元的所述3×3本地滤波运算单元以将所述3×3运算参数及该输入数据进行一3×3卷积运算，以产生多个3×3运算数据；及

一3×3后处理运算程序，驱动所述3×3后处理运算单元以将所述3×3运算数据进行一3×3后处理运算，以产生该3×3后处理运算数据，其中该输出数据为该3×3后处理运算数据。

15.如权利要求14所述的卷积神经网络处理器的数据处理方法，其中该解码模块包含一程序存储器及一指令解码器，且该指令解码步骤包含：

一程序存储子步骤，驱动该程序存储器存储该输入程序；及

一程序解码子步骤，驱动该指令解码器对该输入程序进行解码，以产生该运行指令。

16.如权利要求14所述的卷积神经网络处理器的数据处理方法，其中该平行处理模块包含多个平行子存储器及多个平行子处理器，且该平行处理步骤包含：

一权重参数存储子步骤，驱动所述平行子存储器以平行地存储所述输入权重参数；及

一权重参数处理子步骤，驱动所述平行子处理器，该平行子处理器根据该运行指令，平行地读取所述输入权重参数并进行一运行处理，以产生所述输出权重参数。

17.如权利要求16所述的卷积神经网络处理器的数据处理方法，其中，

当所述输入权重参数为多个非压缩输入权重参数，该运行处理为存储所述非压缩输入权重参数；及

当所述输入权重参数为多个压缩输入权重参数，该运行处理为存储及解压缩所述压缩输入权重参数。

18.如权利要求14所述的卷积神经网络处理器的数据处理方法，其中，

所述输出权重参数还包含一偏压输出权重参数；

该运算模块还包含一偏压分配器；及

该运算步骤还包含：

一偏压运算子步骤，驱动该偏压分配器根据该偏压输出权重参数，以产生多个3×3偏压权重参数，该偏压分配器将所述3×3偏压权重参数提供予该3×3运算子模块。

19.如权利要求14所述的卷积神经网络处理器的数据处理方法，其中，

所述输出权重参数还包含至少一第二输出权重参数；

该运算模块包含一1×1运算子模块；及

该运算步骤还包含：

一第二运算子步骤，驱动该1×1运算子模块接收该3×3后处理运算数据及该至少一第二输出权重参数，以产生一1×1后处理运算数据。

20.如权利要求19所述的卷积神经网络处理器的数据处理方法，其中，

该至少一第二输出权重参数包含多个1×1权重参数；

该1×1运算子模块包含多个1×1卷积分配器组、多个1×1本地卷积运算单元及多个1×1后处理运算单元；及

该第二运算子步骤包含：

一1×1参数分配程序，驱动该至少一1×1卷积分配器组以接收该至少一第二输出权重参数的所述1×1权重参数，并将该至少一第二输出权重参数的所述1×1权重参数分配至所述1×1本地卷积运算单元，其中各该1×1本地卷积运算单元包含一1×1本地暂存器组及一1×1本地滤波运算单元；

一1×1运算参数产生程序，驱动所述1×1本地卷积运算单元的所述1×1本地暂存器组接收该至少一第二输出权重参数的所述1×1权重参数，并根据该至少一第二输出权重参数的所述1×1权重参数产生多个1×1运算参数；

一1×1卷积运算程序，驱动所述1×1本地卷积运算单元的该1×1本地滤波运算单元以将所述1×1运算参数及该3×3后处理运算数据进行一1×1卷积运算，以产生多个1×1运算数据；及

一1×1后处理运算程序，驱动所述1×1后处理运算单元以将所述1×1运算数据进行一1×1后处理运算，以产生该1×1后处理运算数据，其中该输出数据为该1×1后处理运算数据。

21.如权利要求19所述的卷积神经网络处理器的数据处理方法，其中，

所述输出权重参数还包含一偏压输出权重参数；

该运算模块还包含一偏压分配器；及

该运算步骤还包含：

一偏压运算子步骤，驱动该偏压分配器根据该偏压输出权重参数，以产生多个3×3偏压权重参数及多个1×1偏压权重参数，其中该偏压分配器将所述3×3偏压权重参数提供予该3×3运算子模块，并将所述1×1偏压权重参数提供予该1×1运算子模块。

卷积神经网络处理器及其数据处理方法

技术领域

[0001] 本发明涉及一种卷积神经网络处理器及其数据处理方法,且尤其是有关一种具有信息解码单元以及卷积判断单元的卷积神经网络处理器及其数据处理方法。

背景技术

[0002] 卷积神经网络(Convolutional Neural Networks,CNN)近期被广泛的应用于电脑视觉(Computer vision)及影像处理(image processing)领域。然而,近期的应用则多偏重于物体识别及物体检测上,因此,卷积神经网络的硬件设计并没有针对图像处理网络进行优化,因为上述应用并不考虑(1)空间分辨率不会被大量降采样(downsampled)以及(2)模型稀疏性的失效状况(model sparsity),导致极高的内存带宽及极高的运算能力需求。

[0003] 有鉴于此,本发明设计一种可执行高度平行运算的卷积神经网络处理器及其数据处理方法以提供高性能的运算。

发明内容

[0004] 本发明提供的卷积神经网络处理器及其数据处理方法,通过信息解码单元以及卷积判断单元而可执行高度平行运算。

[0005] 依据本发明一实施方式提供一种卷积神经网络处理器,用以运算输入数据,卷积神经网络处理器包含信息解码单元以及卷积判断单元。信息解码单元用以接收输入程序及多个输入权重参数,且包含解码模块及平行处理模块。解码模块接收输入程序,并根据输入程序输出运行指令。平行处理模块与解码模块电性连接,并接收输入权重参数,且平行处理模块包含多个平行处理子模块,平行处理子模块根据运行指令及输入权重参数产生多个输出权重参数。卷积判断单元与信息解码单元电性连接,且包含运算模块。运算模块与平行处理模块电性连接,运算模块依据输入数据与输出权重参数运算而产生输出数据。

[0006] 因此,卷积神经网络处理器可通过信息解码单元的解码模块及平行处理模块,以及卷积判断单元的运算模块执行高度平行运算,进而提供高性能且低功耗的运算。

[0007] 根据前段所述实施方式的卷积神经网络处理器,其中解码模块包含程序存储器及指令解码器。程序存储器存储输入程序。指令解码器与程序存储器电性连接,指令解码器将输入程序解码以输出运行指令。

[0008] 根据前段所述实施方式的卷积神经网络处理器,当输入权重参数为多个非压缩输入权重参数,平行处理子模块包含多个平行子存储器及多个平行子处理器。多个平行子存储器平行地存储非压缩输入权重参数。多个平行子处理器分别与解码模块及平行子存储器电性连接,平行子处理器根据运行指令平行地接收非压缩输入权重参数,产生输出权重参数。

[0009] 根据前段所述实施方式的卷积神经网络处理器,当输入权重参数为多个压缩输入权重参数,平行处理子模块包含多个平行子存储器及多个平行子处理器。多个平行子存储器平行地存储压缩输入权重参数。多个平行子处理器分别与解码模块及平行子存储器电性

连接，平行子处理器根据运行指令平行地接收并解压缩压缩输入权重参数，产生输出权重参数。

[0010] 根据前段所述实施方式的卷积神经网络处理器，其中输入权重参数包含多个第一输入权重参数、输出权重参数包含多个第一输出权重参数。平行处理子模块包含多个平行子存储器及多个平行子处理器。多个平行子存储器平行地存储输入权重参数，平行子存储器包含多个第一平行子存储器。多个第一平行子存储器分别且平行地接收并存储第一输入权重参数。多个平行子处理器分别与解码模块及平行子存储器电性连接，平行子处理器包含多个第一平行子处理器。多个第一平行子处理器分别与第一平行子存储器电性连接，根据运行指令接收第一输入权重参数，以输出第一输出权重参数。

[0011] 根据前段所述实施方式的卷积神经网络处理器，其中第一输出权重参数包含多个3×3权重参数。运算模块包含3×3运算子模块。3×3运算子模块与第一平行子处理器电性连接，并根据第一输出权重参数与输入数据进行运算，以产生3×3后处理运算数据，3×3运算子模块包含多个3×3卷积分配器组、多个3×3本地卷积运算单元及多个3×3后处理运算单元。各3×3卷积分配器组与一第一平行子处理器电性连接，3×3卷积分配器组用以接收及分配第一输出权重参数的3×3权重参数。多个3×3本地卷积运算单元分别与一3×3卷积分配器组电性连接，各3×3本地卷积运算单元包含3×3本地暂存器组及3×3本地滤波运算单元。3×3本地暂存器组与一3×3卷积分配器组电性连接，3×3本地卷积运算单元的3×3本地暂存器组接收并存储第一输出权重参数的3×3权重参数，并根据第一输出权重参数的3×3权重参数，以输出多个3×3运算参数。3×3本地滤波运算单元与3×3本地暂存器组电性连接，3×3本地卷积运算单元的3×3本地滤波运算单元根据3×3运算参数与输入数据进行运算以产生多个3×3运算数据。多个3×3后处理运算单元与3×3本地卷积运算单元电性连接，并依据3×3运算数据进行3×3后处理运算，以产生3×3后处理运算数据，其中输出数据为3×3后处理运算数据。

[0012] 根据前段所述实施方式的卷积神经网络处理器，其中各3×3本地暂存器组包含二子3×3本地暂存器组。二子3×3本地暂存器组交替地存储一3×3权重参数或输出3×3运算参数给3×3本地滤波运算单元。

[0013] 根据前段所述实施方式的卷积神经网络处理器，其中输入权重参数还包含偏压输入权重参数、输出权重参数还包含偏压输出权重参数。平行子存储器还包含偏压平行子存储器。偏压平行子存储器平行地存储偏压输入权重参数。平行子处理器还包含偏压平行子处理器。偏压平行子处理器与偏压平行子存储器电性连接，根据运行指令接收偏压输入权重参数，以输出偏压输出权重参数。

[0014] 根据前段所述实施方式的卷积神经网络处理器，其中偏压输出权重参数包含多个偏压权重参数。运算模块还包含偏压分配器。偏压分配器与偏压平行子处理器、3×3运算子模块电性连接，偏压分配器根据偏压输出权重参数以产生多个3×3偏压权重参数，并将3×3偏压权重参数输出至3×3后处理运算单元。

[0015] 根据前段所述实施方式的卷积神经网络处理器，其中输入权重参数还包含至少一第二输入权重参数、输出权重参数还包含至少一第二输出权重参数。平行子存储器还包含至少一第二平行子存储器。第二平行子存储器分别且平行地接收并存储至少一第二输入权重参数。平行子处理器还包含至少一第二平行子处理器。至少一第二平行子处理器分别与

至少一第二平行子存储器电性连接，根据运行指令接收至少一第二输入权重参数，以输出至少一第二输出权重参数。

[0016] 根据前段所述实施方式的卷积神经网络处理器，其中运算模块包含3×3运算子模块及1×1运算子模块。3×3运算子模块与第一平行子处理器电性连接，并根据第一输出权重参数与输入数据进行运算，以产生3×3后处理运算数据。1×1运算子模块与至少一第二平行子处理器及3×3运算子模块电性连接，并根据至少一第二输出权重参数与3×3后处理运算数据进行运算，以产生1×1后处理运算数据，其中，输出数据可为1×1后处理运算数据。

[0017] 根据前段所述实施方式的卷积神经网络处理器，其中至少一第二输出权重参数包含多个1×1权重参数。1×1运算子模块包含至少一1×1卷积分配器组、多个1×1本地卷积运算单元及多个1×1后处理运算单元。1×1卷积分配器组与至少一第二平行子处理器电性连接，用以接收及分配至少一第二输出权重参数的1×1权重参数。多个1×1本地卷积运算单元与至少一1×1卷积分配器电性连接，各1×1本地卷积运算单元包含1×1本地暂存器组及1×1本地滤波运算单元。1×1本地暂存器组与至少一1×1卷积分配器组电性连接，1×1本地卷积运算单元的1×1本地暂存器组接收并存储第二输出权重参数的1×1权重参数，并根据至少一第二输出权重参数的1×1权重参数，以输出多个1×1运算参数。1×1本地滤波运算单元与1×1本地暂存器组电性连接，1×1本地卷积运算单元的1×1本地滤波运算单元根据1×1运算参数与3×3后处理运算数据进行运算以产生多个1×1运算数据。多个1×1后处理运算单元与1×1本地卷积运算单元电性连接，并依据1×1运算数据进行1×1后处理运算，以产生1×1后处理运算数据。

[0018] 根据前段所述实施方式的卷积神经网络处理器，其中各1×1本地暂存器组包含二子1×1本地暂存器组。二子1×1本地暂存器组交替地存储一1×1权重参数或输出1×1运算参数给1×1运算单元。

[0019] 根据前段所述实施方式的卷积神经网络处理器，其中输入权重参数还包含偏压输入权重参数、输出权重参数还包含偏压输出权重参数。平行子存储器还包含偏压平行子存储器。偏压平行子存储器平行地存储偏压输入权重参数。平行子处理器还包含偏压平行子处理器。偏压平行子处理器与偏压平行子存储器电性连接，根据运行指令接收偏压输入权重参数，以输出偏压输出权重参数。

[0020] 根据前段所述实施方式的卷积神经网络处理器，其中偏压输出权重参数包含多个偏压权重参数。运算模块还包含偏压分配器。偏压分配器与偏压平行子处理器、3×3运算子模块及1×1运算子模块电性连接，偏压分配器根据偏压输出权重参数以产生多个3×3偏压权重参数及多个1×1偏压权重参数。偏压分配器将3×3偏压权重参数输出至3×3后处理运算单元。偏压分配器将1×1偏压权重参数输出至1×1后处理运算单元。

[0021] 依据本发明一实施方式提供一种卷积神经网络处理器的数据处理方法包含接收步骤、指令解码步骤、平行处理步骤及运算步骤。接收步骤驱动信息解码单元接收输入程序及多个输入权重参数，其中信息解码单元包含解码模块及平行处理模块。指令解码步骤驱动解码模块接收输入程序，并根据输入程序，以产生运行指令。平行处理步骤驱动平行处理模块接收输入权重参数，并根据运行指令以平行地处理输入权重参数，以产生多个输出权重参数。运算步骤驱动运算模块接收输入数据及输出权重参数，并根据运行指令以将输入

数据与输出权重参数进行运算，以产生输出数据。

[0022] 因此，卷积神经网络处理器的数据处理方法可通过接收步骤、指令解码步骤、平行处理步骤及运算步骤驱动信息解码单元的解码模块及平行处理模块，以及卷积判断单元的运算模块执行高度平行运算，进而提供高性能且低功耗的运算。

[0023] 根据前段所述实施方式的卷积神经网络处理器的数据处理方法，其中解码模块包含程序存储器及指令解码器，且指令解码步骤包含程序存储子步骤及程序解码子步骤。程序存储子步骤驱动程序存储器存储输入程序。程序解码子步骤驱动指令解码器对输入程序进行解码，以产生运行指令。

[0024] 根据前段所述实施方式的卷积神经网络处理器的数据处理方法，其中平行处理模块包含多个平行子存储器及多个平行子处理器，且平行处理步骤包含权重参数存储子步骤及权重参数处理子步骤。权重参数存储子步骤驱动平行子存储器以平行地存储输入权重参数。权重参数处理子步骤驱动平行子处理器，平行子处理器根据运行指令，平行地读取输入权重参数并进行运行处理，以产生输出权重参数。

[0025] 根据前段所述实施方式的卷积神经网络处理器的数据处理方法，其中当输入权重参数为多个非压缩输入权重参数，运行处理为存储非压缩输入权重参数。当输入权重参数为多个压缩输入权重参数，运行处理为存储及解压缩压缩输入权重参数。

[0026] 根据前段所述实施方式的卷积神经网络处理器的数据处理方法，其中输出权重参数包含多个第一输出权重参数。运算模块包含3×3运算子模块。运算步骤包含第一运算子步骤。第一运算子步骤驱动3×3运算子模块接收输入数据及第一输出权重参数，以产生3×3后处理运算数据。

[0027] 根据前段所述实施方式的卷积神经网络处理器的数据处理方法，其中各第一输出权重参数包含多个3×3权重参数。3×3运算子模块包含多个3×3卷积分配器组、多个3×3本地卷积运算单元及多个3×3后处理运算单元。第一运算子步骤包含3×3参数分配程序、3×3运算参数产生程序、3×3卷积运算程序及3×3后处理运算程序。3×3参数分配程序驱动3×3卷积分配器组接收第一输出权重参数的3×3权重参数，并将第一输出权重参数的3×3权重参数分配至3×3本地卷积运算单元。各3×3本地卷积运算单元包含3×3本地暂存器组及3×3本地滤波运算单元。3×3运算参数产生程序驱动3×3本地卷积运算单元的3×3本地暂存器组接收第一输出权重参数的3×3权重参数，并根据第一输出权重参数的3×3权重参数产生多个3×3运算参数。3×3卷积运算程序驱动3×3本地卷积运算单元的3×3本地滤波运算单元以将3×3运算参数及输入数据进行3×3卷积运算，以产生多个3×3运算数据。3×3后处理运算程序驱动3×3后处理运算单元以将3×3运算数据进行3×3后处理运算，以产生3×3后处理运算数据，其中输出数据为3×3后处理运算数据。

[0028] 根据前段所述实施方式的卷积神经网络处理器的数据处理方法，其中输出权重参数还包含偏压输出权重参数。运算模块还包含偏压分配器。运算步骤还包含偏压运算子步骤。偏压运算子步骤驱动偏压分配器根据偏压输出权重参数，以产生多个3×3偏压权重参数，偏压分配器将3×3偏压权重参数提供予3×3运算子模块。

[0029] 根据前段所述实施方式的卷积神经网络处理器的数据处理方法，其中输出权重参数还包含至少一第二输出权重参数。运算模块包含1×1运算子模块。运算步骤还包含第二运算子步骤。第二运算子步骤驱动1×1运算子模块接收3×3后处理运算数据及至少一第二

输出权重参数，以产生1×1后处理运算数据。

[0030] 根据前段所述实施方式的卷积神经网络处理器的数据处理方法，其中至少一第二输出权重参数包含多个1×1权重参数。1×1运算子模块包含多个1×1卷积分配器组、多个1×1本地卷积运算单元及多个1×1后处理运算单元。第二运算子步骤包含1×1参数分配程序、1×1运算参数产生程序、1×1卷积运算程序及1×1后处理运算程序。1×1参数分配程序驱动1×1卷积分配器组以接收至少一第二输出权重参数的1×1权重参数，并将至少一第二输出权重参数的1×1权重参数分配至1×1本地卷积运算单元，其中各1×1本地卷积运算单元包含1×1本地暂存器组及1×1本地滤波运算单元。1×1运算参数产生程序驱动1×1本地卷积运算单元的1×1本地暂存器组接收至少一第二输出权重参数的1×1权重参数，并根据至少一第二输出权重参数的1×1权重参数产生多个1×1运算参数。1×1卷积运算程序驱动1×1本地卷积运算单元的1×1本地滤波运算单元以将1×1运算参数及3×3后处理运算数据进行1×1卷积运算，以产生多个1×1运算数据。1×1后处理运算程序驱动1×1后处理运算单元以将1×1运算数据进行1×1后处理运算，以产生1×1后处理运算数据，其中输出数据为1×1后处理运算数据。

[0031] 根据前段所述实施方式的卷积神经网络处理器的数据处理方法，其中输出权重参数还包含偏压输出权重参数。运算模块还包含偏压分配器。运算步骤还包含偏压运算子步骤。偏压运算子步骤驱动偏压分配器根据偏压输出权重参数，以产生多个3×3偏压权重参数及多个1×1偏压权重参数，其中偏压分配器将3×3偏压权重参数提供予3×3运算子模块，并将1×1偏压权重参数提供予1×1运算子模块。

附图说明

[0032] 图1示出依照本发明的一结构态样的一实施方式的卷积神经网络处理器的方框图；

[0033] 图2示出依照本发明的另一结构态样的一实施方式的卷积神经网络处理器的方框图；

[0034] 图3示出依照图2的结构态样的实施方式的卷积神经网络处理器的3×3运算子模块的方框图；

[0035] 图4示出依照图3的结构态样的实施方式的卷积神经网络处理器的3×3运算子模块的3×3本地卷积运算单元示意图；

[0036] 图5示出依照本发明的又一结构态样的一实施方式的卷积神经网络处理器的方框图；

[0037] 图6示出依照图5的结构态样的实施方式的卷积神经网络处理器的1×1运算子模块的方框图；

[0038] 图7示出依照图6的结构态样的实施方式的卷积神经网络处理器的1×1运算子模块的1×1本地卷积运算单元示意图；

[0039] 图8示出依照本发明的一方法态样的一实施方式的卷积神经网络处理器的数据处理方法的步骤方框图；

[0040] 图9示出依照图8的方法态样的实施方式的卷积神经网络处理器的数据处理方法的指令解码步骤的步骤方框图；

[0041] 图10示出依照图8的方法态样的实施方式的卷积神经网络处理器的数据处理方法的平行处理步骤的步骤方框图；

[0042] 图11示出依照图8的方法态样的实施方式的卷积神经网络处理器的数据处理方法的运算步骤的步骤方框图;以及

[0043] 图12示出依照图8的方法态样的另一实施方式的卷积神经网络处理器的数据处理方法的运算步骤的步骤方框图。

[0044] 附图标记说明：

[0045] 100:卷积神经网络处理器

[0046] 102:输入程序

[0047] 104:输入权重参数

[0048] 106:输入数据

[0049] 1062:3×3后处理运算数据

[0050] 1064:1×1后处理运算数据

[0051] 108:输出数据

110:信息解码单元

[0052] 111:解码模块

[0053] 1111:程序存储器

[0054] 1112:指令解码器

[0055] 112:平行处理模块

[0056] 1121:平行处理子模块

[0057] 1121a:平行子存储器

[0058] 1121aa:第一平行子存储器

[0059] 1121ab:偏压平行子存储器

[0060] 1121ac:第二平行子存储器

[0061] 1121b:平行子处理器

[0062] 1121ba:第一平行子处理器

[0063] 1121bb:偏压平行子处理器序

[0064] 1121bc:第二平行子处理器

[0065] 120:卷积判断单元

[0066] 121:运算模块

[0067] 1211:3×3运算子模块

[0068] 1211a:3×3运算电路

[0069] 1211b:3×3本地卷积运算单元

[0070] 1211c:3×3本地暂存器组

[0071] 1211ca、1211cb:子3×3本地

[0072] 1211d:3×3本地滤波运算单

[0073] 元

[0074] 1211e:3×3后处理运算单元

[0075] 1211f:第一3×3卷积分配器

1213d:1×1本地滤波运算单元

1213e:1×1后处理运算单元

1213f:第一1×1卷积分配器

1213g:第二1×1卷积分配器

122:控制器

s200:卷积神经网络处理器的数据处理方法

s210:接收步骤

s220:指令解码步骤

s221:程序存储子步骤

s222:程序解码子步骤

s230:平行处理步骤

s231:权重参数存储子步骤

s232:权重参数处理子步骤

s240:运算步骤

s241:第一运算子步骤

s2411:3×3参数分配程序

s2412:3×3运算参数产生程

s2413:3×3卷积运算程序

s2414:3×3后处理运算程序

s242:偏压运算子步骤

s243:第二运算子步骤

s2431:1×1参数分配程序

s2432:1×1运算参数产生程序

s2433:1×1卷积运算程序

s2434:1×1后处理运算程序暂存器组

[0076] 1211g:第二3×3卷积分配器

[0077] 1212:偏压分配器

[0078] 1213:1×1运算子模块

[0079] 1213a:1×1运算电路

[0080] 1213b:1×1本地卷积运算单

[0081] 元

[0082] 1213c:1×1本地暂存器组

[0083] 1213ca、1213cb:子1×1本地

[0084] 暂存器组

具体实施方式

[0085] 以下将参照附图说明本发明的多个实施例。为明确说明起见,许多实务上的细节将在以下叙述中一并说明。然而,应了解到,这些实务上的细节不应用以限制本发明。也就是说,在本发明部分实施例中,这些实务上的细节是非必要的。此外,为简化附图起见,一些现有惯用的结构与元件在附图中将以简单示意的方式示出的;并且重复的元件将可能使用相同的编号表示的。

[0086] 图1示出依照本发明一结构态样的一态样的卷积神经网络处理器100的方框图。由图1可知,卷积神经网络处理器100包含信息解码单元110及卷积判断单元120。卷积判断单元120与信息解码单元110电性连接。

[0087] 信息解码单元110接收输入程序102及多个输入权重参数104。信息解码单元110包含解码模块111及平行处理模块112。解码模块111接收输入程序102,并根据输入程序102输出运行指令。平行处理模块112与解码模块111电性连接,平行处理模块112接收输入权重参数104与运行指令。平行处理模块112包含多个平行处理子模块1121,平行处理子模块1121根据运行指令及输入权重参数104产生多个输出权重参数。卷积判断单元120包含运算模块121。运算模块121与平行处理模块112电性连接,运算模块121依据输入数据106与输出权重参数运算而产生输出数据108。详细来说,卷积神经网络处理器100的信息解码单元110于接收输入程序102及输入权重参数104后,利用解码模块111产生运行指令以处理输入权重参数104。平行处理模块112的各平行处理子模块1121可分别与解码模块111电性连接,以分别根据运行指令产生输出权重参数。运算模块121可根据输入数据106及平行处理模块112所产生的输出权重参数进行运算以产生输出数据108。输入数据106可为存储于区域缓冲区(block buffer bank)中的数据或是来自外部的数据。此外,卷积神经网络处理器100可利用区域缓冲区代替输入缓冲区及输出缓冲区以节省外部存储器的带宽。因此,卷积神经网络处理器100可通过信息解码单元110及卷积判断单元120执行高度平行运算以提供高性能的运算。

[0088] 解码模块111可包含程序存储器1111及指令解码器1112。程序存储器1111可存储输入程序102。指令解码器1112与程序存储器1111电性连接。指令解码器1112将输入程序102解码以输出运行指令。也就是说,解码模块111于接收输入程序102后,将输入程序102存储于程序存储器1111中并通过指令解码器1112进行解码以产生运行指令,进而通过运行指令驱动各平行处理子模块1121处理输入权重参数104以产生输出权重参数。

[0089] 当输入权重参数104为非压缩输入权重参数时，多个平行处理子模块1121包含多个平行子存储器1121a及多个平行子处理器1121b。平行子存储器1121a平行地存储非压缩输入权重参数。平行子处理器1121b分别与解码模块111及一平行子存储器1121a电性连接。平行子处理器1121b根据运行指令平行地接收非压缩输入权重参数以产生输出权重参数。详细来说，各平行处理子模块1121分别可包含一平行子存储器1121a及一平行子处理器1121b。平行处理模块112于接收输入权重参数104后，将输入权重参数104分别且平行地存储于各平行处理子模块1121的平行子存储器1121a中。由于各平行处理子模块1121分别与解码模块111电性连接，因此，各平行子处理器1121b可分别根据运行指令平行地从平行子存储器1121a接收非压缩输入权重参数以产生输出权重参数。因此，平行处理模块112可平行地处理输入权重参数104以产生输出权重参数。

[0090] 当输入权重参数104为多个压缩输入权重参数时，多个平行处理子模块1121包含多个平行子存储器1121a及多个平行子处理器1121b。平行子存储器1121a平行地存储压缩输入权重参数。平行子处理器1121b分别与解码模块111及一平行子存储器1121a电性连接。平行子处理器1121b根据运行指令平行地接收并解压缩此些压缩输入权重参数以产生输出权重参数。详细来说，各平行处理子模块1121分别可包含一平行子存储器1121a及一平行子处理器1121b。平行处理模块112于接收输入权重参数104后，将输入权重参数104分别且平行地存储于各平行处理子模块1121的平行子存储器1121a中。由于各平行处理子模块1121分别与解码模块111电性连接，因此，各平行子处理器1121b可分别根据运行指令平行地从平行子存储器1121a接收压缩输入权重参数，并将压缩输入权重参数进行解码以产生输出权重参数。因此，平行处理模块112可平行地处理输入权重参数104以产生输出权重参数。

[0091] 请配合参照图1、图2、图3及图4。图2示出依照本发明另一结构态样的一实施方式的卷积神经网络处理器100的方框图。图3示出依照图2结构态样的实施方式的卷积神经网络处理器100的3×3运算子模块1211的方框图。图4示出依照图3结构态样的实施方式的卷积神经网络处理器100的3×3运算子模块1211的3×3本地卷积运算单元1211b的示意图。在图2至图4的实施方式中，输入权重参数104可包含多个第一输入权重参数及偏压输入权重参数。输出权重参数包含多个第一输出权重参数及偏压输出权重参数。平行处理子模块1121包含多个平行子存储器1121a及多个平行子处理器1121b。平行子存储器1121a平行地存储输入权重参数104，并包含多个第一平行子存储器1121aa及偏压平行子存储器1121ab。第一平行子存储器1121aa分别且平行地接收并存储第一输入权重参数。偏压平行子存储器1121ab平行地存储偏压输入权重参数。平行子处理器1121b分别与解码模块111及平行子存储器1121a电性连接，并包含多个第一平行子处理器1121ba及偏压平行子处理器1121bb。第一平行子处理器1121ba分别与第一平行子存储器1121aa中的一者电性连接，并根据运行指令接收第一输入权重参数，以输出第一输出权重参数。偏压平行子处理器1121bb与偏压平行子存储器1121ab电性连接，并根据运行指令接收偏压输入权重参数，以输出偏压输出权重参数。在图2的实施方式中，第一平行子存储器1121aa及第一平行子处理器1121ba的数量均为9，然在其他实施方式中，第一平行子存储器1121aa及第一平行子处理器1121ba的数量可为9的倍数，本公开不以此为限。偏压平行子存储器1121ab及偏压平行子处理器1121bb的数量均为1，但本发明不以此为限。详细来说，平行处理模块112于接收输入权重参数104后，将输入权重参数104中的第一输入权重参数存储于第一平行子存储器1121aa中，以及将偏

压输入权重参数存储于偏压平行子存储器1121ab中。第一平行子处理器1121ba根据运行指令从第一平行子存储器1121aa中读取第一输入权重参数，并进行处理以产生第一输出权重参数。偏压平行子处理器1121bb根据运行指令从偏压平行子存储器1121ab中读取偏压输入权重参数，并进行处理以产生偏压输出权重参数。

[0092] 各第一输出权重参数包含多个3×3权重参数。运算模块121可包含3×3运算子模块1211及偏压分配器1212。3×3运算子模块1211与第一平行子处理器1121ba电性连接，并根据第一输出权重参数与输入数据106进行运算，以产生3×3后处理运算数据1062。3×3运算子模块1211包含3×3卷积分配器组、3×3本地卷积运算单元1211b及3×3后处理运算单元1211e。各3×3卷积分配器组与一第一平行子处理器1121ba电性连接，3×3卷积分配器组用以接收及分配第一输出权重参数的3×3权重参数。3×3本地卷积运算单元1211b分别与一3×3卷积分配器组电性连接，并包含3×3本地暂存器组1211c及3×3本地滤波运算单元1211d。3×3本地暂存器组1211c与一3×3卷积分配器组电性连接，3×3本地卷积运算单元1211b的3×3本地暂存器组1211c接收并存储第一输出权重参数的些3×3权重参数，并根据第一输出权重参数的3×3权重参数以输出多个3×3运算参数。3×3本地滤波运算单元1211d与3×3本地暂存器组1211c电性连接，3×3本地卷积运算单元1211b的3×3本地滤波运算单元1211d根据3×3运算参数与输入数据106进行运算以产生多个3×3运算数据。详细来说，3×3本地滤波运算单元1211d可执行3×3卷积运算，当第一平行子处理器1121ba的数量为9时，3×3本地滤波运算单元1211d的空间滤波位置(spatial filter position)分别可对应第一平行子处理器1121ba；当第一平行子处理器1121ba的数量为18时，3×3本地滤波运算单元1211d的空间滤波位置可对应二第一平行子处理器1121ba，以此类推，本公开不另赘述。3×3后处理运算单元1211e与3×3本地卷积运算单元1211b电性连接，并依据3×3运算数据进行3×3后处理运算以产生3×3后处理运算数据1062。卷积神经网络处理器100的输出数据108可为3×3后处理运算数据1062。偏压分配器1212与偏压平行子处理器1121bb、3×3运算子模块1211电性连接。偏压分配器1212根据偏压输出权重参数以产生多个3×3偏压权重参数，并将3×3偏压权重参数输出至3×3后处理运算单元1211e。

[0093] 在图3中，3×3运算子模块1211包含多个3×3运算电路1211a，3×3运算电路1211a的数量可为32。各3×3运算电路1211a是由多个3×3本地卷积运算单元1211b及一3×3后处理运算单元1211e所组成，3×3本地卷积运算单元1211b的数量可为32。也就是说，3×3运算子模块1211中的3×3本地卷积运算单元1211b的数量为1024，3×3后处理运算单元1211e的数量为32。

[0094] 请配合参照图3及图4，3×3运算子模块1211于接收第一输出权重参数的3×3权重参数后可通过3×3卷积分配器组将3×3权重参数分配至3×3本地卷积运算单元1211b。在图4中，3×3卷积分配器组的配置是采用二阶段分配法，3×3卷积分配器组包含第一3×3卷积分配器1211f及多个第二3×3卷积分配器1211g。第一3×3卷积分配器1211f与第一平行子处理器1121ba电性连接以接收并分配第一输出权重参数的3×3权重参数至第二3×3卷积分配器1211g，第二3×3卷积分配器1211g于接收3×3权重参数后，将3×3权重参数分配至3×3本地卷积运算单元1211b，本发明虽是利用二阶段的分配方法，然其分配方式并不以此为限。3×3本地暂存器组1211c可包含二子3×3本地暂存器组1211ca、1211cb。二子3×3本地暂存器组1211ca、1211cb结合一个多工器可交替地存储3×3权重参数或输出3×3运算

参数给3×3本地滤波运算单元1211d。也就是说，当子3×3本地暂存器组1211ca用以存储3×3权重参数时，子3×3本地暂存器组1211cb输出3×3运算参数给3×3本地滤波运算单元1211d；当子3×3本地暂存器组1211cb用以存储3×3权重参数时，子3×3本地暂存器组1211ca输出3×3运算参数给3×3本地滤波运算单元1211d，即本公开的3×3本地暂存器组1211c是以乒乓(ping-pong)的方式执行存储3×3权重参数及输出3×3运算参数。

[0095] 3×3本地滤波运算单元1211d可根据3×3运算参数及输入数据106进行3×3卷积运算以产生3×3运算数据。举例来说，输入数据106的图块大小可为6×4，3×3本地滤波运算单元1211d可根据3×3运算参数与输入数据106进行3×3卷积运算。为了实现高度平行的运算，卷积神经网络处理器100可在3×3运算子模块1211中布署多个乘法器，于3×3本地滤波运算单元1211d中的乘法器的数量可为73728。3×3后处理运算单元1211e于接收3×3本地滤波运算单元1211d所产生的3×3运算数据及偏压分配器所产生的3×3偏压权重参数后，可根据3×3运算数据及3×3偏压权重参数进行3×3后处理运算以产生3×3后处理运算数据1062。在图3及图4的实施方式中，3×3后处理运算数据1062即为卷积神经网络处理器100的输出数据108。

[0096] 在图2中，卷积判断单元120还包含控制器122。控制器122与信息解码单元110电性连接。详细来说，控制器122与指令解码器1112电性连接以接收运行指令，并根据运行指令控制运算模块121的3×3运算子模块1211及偏压分配器1212。

[0097] 图5示出依照本发明的又一结构态样的一实施方式的卷积神经网络处理器100的方框图。图6示出依照图5的结构态样的实施方式的卷积神经网络处理器100的1×1运算子模块1213的方框图。图7示出依照图6的结构态样的实施方式的卷积神经网络处理器100的1×1运算子模块1213的1×1本地卷积运算单元1213b示意图。图5的卷积神经网络处理器100与图2的卷积神经网络处理器100的差异在于，图5的卷积神经网络处理器100的平行子存储器1121a还包含至少一第二平行子存储器1121ac，平行子处理器1121b还包含至少一第二平行子处理器1121bc，及运算模块121还包含1×1运算子模块1213。此外，输入权重参数104还包含至少一第二输入权重参数。输出权重参数还包含至少一第二输出权重参数。至少一第二平行子存储器1121ac分别且平行地接收并存储至少一第二输入权重参数。至少一第二平行子处理器1121bc分别与至少一第二平行子存储器1121ac电性连接，并根据运行指令接收至少一第二输入权重参数以输出至少一第二输出权重参数。3×3运算子模块1211的配置与图2的卷积神经网络处理器100中的3×3运算子模块1211相同，在此不另赘述。在图5的实施方式中，第一平行子存储器1121aa及第一平行子处理器1121ba的数量均为9，第二平行子存储器1121ac及第二平行子处理器1121bc的数量均为1，然在其他实施方式中，当第一平行子存储器1121aa及第一平行子处理器1121ba的数量为18时，第二平行子存储器1121ac及第二平行子处理器1121bc的数量均为2，以此类推，本公开不以此为限。偏压平行子存储器1121ab及偏压平行子处理器1121bb的数量均为1，但本发明不以此为限。

[0098] 详细来说，平行处理模块112于接收输入权重参数104后，将输入权重参数104中的第一输入权重参数存储于第一平行子存储器1121aa中，将输入权重参数104中的第二输入权重参数存储于第二平行子存储器1121ac中，以及将偏压输入权重参数存储于偏压平行子存储器1121ab中。图5的第一平行子处理器1121ba及偏压平行子处理器1121bb的运行方式与图2的第一平行子处理器1121ba及偏压平行子处理器1121bb相同，在此不另赘述。第二平

行子处理器1121bc根据运行指令从第二平行子存储器1121ac中读取第二输入权重参数，并进行处理以产生第二输出权重参数。

[0099] 1×1运算子模块1213与至少一第二平行子处理器1121bc及3×3运算子模块1211电性连接，并根据至少一第二输出权重参数与3×3后处理运算数据1062进行运算以产生1×1后处理运算数据1064。1×1运算子模块1213包含至少一1×1卷积分配器组、多个1×1本地卷积运算单元及多个1×1后处理运算单元1213e。至少一1×1卷积分配器组与至少一第二平行子处理器1121bc电性连接，用以接收及分配至少一第二输出权重参数的1×1权重参数。1×1本地卷积运算单元1213b与至少一1×1卷积分配器电性连接。各1×1本地卷积运算单元1213b包含1×1本地暂存器组1213c及1×1本地滤波运算单元1213d。1×1本地暂存器组1213c与至少一1×1卷积分配器组电性连接。1×1本地卷积运算单元1213b的1×1本地暂存器组1213c接收并存储至少一第二输出权重参数的1×1权重参数，并根据至少一第二输出权重参数的1×1权重参数，以输出1×1运算参数。1×1本地滤波运算单元1213d与1×1本地暂存器组1213c电性连接。1×1本地卷积运算单元1213b的1×1本地滤波运算单元1213d根据1×1运算参数与3×3后处理运算数据1062进行运算以产生多个1×1运算数据。详细来说，1×1本地滤波运算单元1213d可执行1×1卷积运算，当第二平行子处理器1121bc的数量为1时，1×1本地滤波运算单元1213d的空间滤波位置可对应第二平行子处理器1121bc；当第二平行子处理器1121bc的数量为2时，1×1本地滤波运算单元1213d的空间滤波位置可对应二第二平行子处理器1121bc，以此类推，本公开不另赘述。1×1后处理运算单元1213e与1×1本地卷积运算单元1213b电性连接，并依据1×1运算数据进行1×1后处理运算以产生1×1后处理运算数据1064。卷积神经网络处理器100的输出数据108为1×1后处理运算数据1064。图5的偏压平行子存储器1121ab及偏压平行子处理器1121bb与图2的偏压平行子存储器1121ab及偏压平行子处理器1121bb相同，在此不另赘述。图5的偏压分配器1212与3×3运算子模块1211的配置关系与图2的偏压分配器1212与3×3运算子模块1211的配置关系相同，在此不另赘述。

[0100] 详细来说，图5的偏压分配器1212与偏压平行子处理器1121bb、3×3运算子模块1211及1×1运算子模块1213电性连接。偏压分配器1212根据偏压输出权重参数以产生多个3×3偏压权重参数及多个1×1偏压权重参数。偏压分配器1212将3×3偏压权重参数输出至3×3后处理运算单元1211e。偏压分配器1212将1×1偏压权重参数输出至1×1后处理运算单元1213e。

[0101] 在图6中，1×1运算子模块1213包含多个1×1运算电路1213a，1×1运算电路1213a的数量可为32。各1×1运算电路1213a是由多个1×1本地卷积运算单元1213b及一1×1后处理运算单元1213e所组成，1×1本地卷积运算单元1213b的数量可为32。也就是说，1×1运算子模块1213中的1×1本地卷积运算单元1213b的数量为1024，1×1后处理运算单元1213e的数量为32。

[0102] 请配合参照图6及图7，1×1运算子模块1213于接收第二输出权重参数的1×1权重参数后可通过1×1卷积分配器组将1×1权重参数分配至1×1本地卷积运算单元1213b。在图7中，1×1卷积分配器组的配置是采用二阶段分配法，并包含第一1×1卷积分配器1213f及多个第二1×1卷积分配器1213g，其作动方式与3×3卷积分配器组相同，在此不另赘述。1×1本地暂存器组1213c可包含二子1×1本地暂存器组1213ca、1213cb。二子1×1本地暂存

器组1213ca、1213cb结合一个多工器可交替地存储1×1权重参数或输出1×1运算参数给1×1本地滤波运算单元1213d。1×1本地暂存器组1213c的作动方式与3×3本地暂存器组1211c相同，在此不另赘述。也就是说，本公开的3×3本地暂存器组1211c及1×1本地暂存器组1213c皆是以乒乓(ping-pong)的方式作动。因此，1×1本地滤波运算单元1213d可根据1×1运算参数及3×3后处理运算数据1062进行1×1后处理运算以产生1×1运算数据。在图5至图7的实施方式中，1×1后处理运算数据1064即为卷积神经网络处理器100的输出数据108。

[0103] 为了实现高度平行的运算，卷积神经网络处理器100可在3×3运算子模块1211及1×1运算子模块1213中布署多个乘法器，举例来说，3×3本地滤波运算单元1211d中的乘法器的数量可为73728，1×1本地滤波运算单元1213d中的乘法器的数量可为8192。此外，图5中的控制器122与图2中的控制器122相同，在此不另赘述。

[0104] 图8示出依照本发明的一方法态样的一实施方式的卷积神经网络处理器的数据处理方法s200的步骤方框图。在图8中，卷积神经网络处理器的数据处理方法s200包含接收步骤s210、指令解码步骤s220、平行处理步骤s230及运算步骤s240。

[0105] 请配合参照图1，详细来说，接收步骤s210驱动信息解码单元110接收输入程序102及多个输入权重参数104。信息解码单元110包含解码模块111及平行处理模块112。指令解码步骤s220驱动解码模块111接收输入程序102，并根据输入程序102产生运行指令。平行处理步骤s230驱动平行处理模块112接收输入权重参数104，并根据运行指令以平行地处理输入权重参数104以产生多个输出权重参数。运算步骤s240驱动运算模块121接收输入数据106及输出权重参数，并根据运行指令以将输入数据106与输出权重参数进行运算以产生输出数据108。也就是说，卷积神经网络处理器100的信息解码单元110可通过接收步骤s210接收输入程序102及输入权重参数104以执行指令解码步骤s220及平行处理步骤s230。由于平行处理模块112与解码模块111电性连接，因此，平行处理模块112可根据解码模块111于指令解码步骤s220中所产生的运行指令产生输出权重参数，即平行处理步骤s230。此外，运算模块121与平行处理模块112电性连接，因此，于运算步骤s240中，运算模块121可于接收输入数据106及输出权重参数后，根据输入数据106及输出权重参数进行运算以产生输出数据108。因此，卷积神经网络处理器的数据处理方法s200可通过接收步骤s210、指令解码步骤s220、平行处理步骤s230及运算步骤s240驱动信息解码单元110的解码模块111及平行处理模块112，以及卷积判断单元120的运算模块121执行高度平行运算，进而提供高性能且低功耗的运算。

[0106] 举例来说，在图8中，卷积神经网络处理器的数据处理方法s200的接收步骤s210所接收的输入程序102及输入权重参数104可包含对应多个输入数据106的相关指令及参数。于执行指令解码步骤s220及平行处理步骤s230时，将对应多个输入数据106的相关指令及参数存储于程序存储器1111及平行子存储器1121a中。于执行指令解码步骤s220及平行处理步骤s230时，可针对与其中一输入数据106的相关指令及参数进行处理，以于执行运算步骤s240时，针对所述其中一输入数据106进行运算，并于执行运算步骤s240期间，卷积神经网络处理器的数据处理方法s200可针对其中的另一输入数据106的相关指令及参数进行处理，即针对所述其中的另一输入数据106执行指令解码步骤s220以及平行处理步骤s230。换言之，卷积神经网络处理器的数据处理方法s200是先将全部的输入数据106的相关指令及

参数都存储于程序存储器1111及平行子存储器1121a中，然后再执行每一个输入数据106所对应的指令解码步骤s220、平行处理步骤s230及运算步骤s240。此外，当运算步骤s240在针对所述其中一输入数据106进行运算时，指令解码步骤s220以及平行处理步骤s230可针对所述其中的另一输入数据106的相关指令及参数进行处理。因此，卷积神经网络处理器的数据处理方法s200于执行接收步骤s210后可对多个输入数据106各别进行运算。

[0107] 图9示出依照图8的方法态样的实施方式的卷积神经网络处理器的数据处理方法s200的指令解码步骤s220的步骤方框图。解码模块111可包含程序存储器1111及指令解码器1112。指令解码步骤s220可包含程序存储子步骤s221及程序解码子步骤s222。程序存储子步骤s221驱动程序存储器1111存储输入程序102。程序解码子步骤s222驱动指令解码器1112对输入程序102进行解码以产生运行指令。也就是说，卷积神经网络处理器100可通过程序存储子步骤s221及程序解码子步骤s222驱动解码模块111接收输入程序102，并将输入程序102存储于程序存储器1111中，再通过指令解码器1112对存储于程序存储器1111中的输入程序102进行解码以产生运行指令。

[0108] 图10示出依照图8的方法态样的实施方式的卷积神经网络处理器的数据处理方法s200的平行处理步骤s230的步骤方框图。平行处理模块112可包含多个平行子存储器1121a及多个平行子处理器1121b。平行处理步骤s230包含权重参数存储子步骤s231及权重参数处理子步骤s232。权重参数存储子步骤s231驱动平行子存储器1121a以平行地存储输入权重参数104。权重参数处理子步骤s232驱动平行子处理器1121b。平行子处理器1121b根据运行指令平行地读取输入权重参数104并进行运行处理以产生输出权重参数。也就是说，卷积神经网络处理器100可通过权重参数存储子步骤s231及权重参数处理子步骤s232驱动平行处理模块112接收输入权重参数104，并将输入权重参数104存储于平行子存储器1121a中，平行子处理器1121b再根据运行指令对存储于平行子存储器1121a中的输入权重参数104进行运行处理以产生输出权重参数。当输入权重参数104为非压缩输入权重参数时，运行处理可为存储非压缩输入权重参数。当输入权重参数104为压缩输入权重参数时，运行处理可为存储及解压缩压缩输入权重参数。

[0109] 图11示出依照图8的方法态样的实施方式的卷积神经网络处理器的数据处理方法s200的运算步骤s240的步骤方框图。请配合参照图2至图4。输出权重参数可包含多个第一输出权重参数及偏压输出权重参数。第一输出权重参数包含多个3×3权重参数。运算模块121可包含3×3运算子模块1211及偏压分配器1212。3×3运算子模块1211包含多个3×3卷积分配器组、多个3×3本地卷积运算单元1211b及多个3×3后处理运算单元1211e。运算步骤s240可包含第一运算子步骤s241及偏压运算子步骤s242。第一运算子步骤s241包含3×3参数分配程序s2411、3×3运算参数产生程序s2412、3×3卷积运算程序s2413及3×3后处理运算程序s2414。3×3参数分配程序s2411驱动3×3卷积分配器组接收第一输出权重参数的3×3权重参数，并将第一输出权重参数的3×3权重参数分配至3×3本地卷积运算单元1211b，其中各3×3本地卷积运算单元1211b包含3×3本地暂存器组1211c及3×3本地滤波运算单元1211d。3×3运算参数产生程序s2412驱动3×3本地卷积运算单元1211b的3×3本地暂存器组1211c接收第一输出权重参数的3×3权重参数，并根据第一输出权重参数的3×3权重参数产生多个3×3运算参数。3×3卷积运算程序s2413驱动3×3本地卷积运算单元1211b的3×3本地滤波运算单元1211d以将3×3运算参数及输入数据106进行3×3卷积运算

以产生多个3×3运算数据。3×3后处理运算程序s2414驱动3×3后处理运算单元1211e以将3×3运算数据进行3×3后处理运算以产生3×3后处理运算数据1062。偏压运算子步骤s242驱动偏压分配器1212根据偏压输出权重参数以产生多个3×3偏压权重参数。偏压分配器1212将3×3偏压权重参数提供予3×3运算子模块1211。也就是说，卷积神经网络处理器100可通过第一运算子步骤s241及偏压运算子步骤s242产生3×3后处理运算数据1062。详细来说，3×3运算子模块1211可用以执行第一运算子步骤s241，3×3运算子模块1211的3×3卷积分配器组可执行3×3参数分配程序s2411以将3×3权重参数分配至不同的3×3本地卷积运算单元1211b中的3×3本地暂存器组1211c，以利3×3本地暂存器组1211c执行3×3运算参数产生程序s2412。3×3本地暂存器组1211c可包含二子3×3本地暂存器组1211ca、1211cb。二子3×3本地暂存器组1211ca、1211cb以乒乓的方式作动，进而接收3×3权重参数并输出3×3运算参数至3×3本地滤波运算单元1211d。3×3本地滤波运算单元1211d于3×3卷积运算程序s2413中根据3×3运算参数及输入数据106进行3×3卷积运算以产生3×3运算数据。于3×3后处理运算程序s2414中，3×3后处理运算单元1211e根据偏压分配器1212于偏压运算子步骤s242中所输出的3×3偏压权重参数及3×3运算数据执行3×3后处理运算以产生3×3后处理运算数据1062。在图2至图4及图11的实施方式中，3×3后处理运算数据1062可为卷积神经网络处理器100的输出数据108。

[0110] 图12示出依照图8的方法态样的另一实施方式的卷积神经网络处理器的数据处理方法s200的运算步骤s240的步骤方框图。请配合参照图5至图7。输出权重参数可包含多个第一输出权重参数、至少一第二输出权重参数及偏压输出权重参数。第一输出权重参数包含多个3×3权重参数。至少一第二输出权重参数包含多个1×1权重参数。运算模块121可包含3×3运算子模块1211、1×1运算子模块1213及偏压分配器1212。3×3运算子模块1211包含多个3×3卷积分配器组、多个3×3本地卷积运算单元1211b及多个3×3后处理运算单元1211e。1×1运算子模块包含多个1×1卷积分配器组、多个1×1本地卷积运算单元1213b及多个1×1后处理运算单元1213e。运算步骤s240可包含第一运算子步骤s241、第二运算子步骤s243及偏压运算子步骤s242。图12的第一运算子步骤s241与图11的第一运算子步骤s241相同，在此不另赘述。第二运算子步骤s243驱动1×1运算子模块1213接收3×3后处理运算数据1062及至少一第二输出权重参数以产生1×1后处理运算数据1064。第二运算子步骤s243包含1×1参数分配程序s2431、1×1运算参数产生程序s2432、1×1卷积运算程序s2433及1×1后处理运算程序s2434。1×1参数分配程序s2431驱动至少一1×1卷积分配器组以接收至少一第二输出权重参数的1×1权重参数，并将至少一第二输出权重参数的1×1权重参数分配至1×1本地卷积运算单元s1213b，其中各1×1本地卷积运算单元s1213b包含1×1本地暂存器组s1213c及1×1本地滤波运算单元s1213d。1×1运算参数产生程序s2432驱动1×1本地卷积运算单元1213b的1×1本地暂存器组1213c接收至少一第二输出权重参数的1×1权重参数，并根据至少一第二输出权重参数的1×1权重参数产生多个1×1运算参数。1×1卷积运算程序s2433驱动1×1本地卷积运算单元1213b的1×1本地滤波运算单元1213d以将1×1运算参数及3×3后处理运算数据1062进行1×1卷积运算以产生多个1×1运算数据。1×1后处理运算程序s2434驱动1×1后处理运算单元1213e以将1×1运算数据进行1×1后处理运算以产生1×1后处理运算数据1064。也就是说，卷积神经网络处理器100可通过第一运算子步骤s241、第二运算子步骤s243及偏压运算子步骤s242产生1×1后处理运算数据

1064。详细来说，1×1运算子模块1213可用以执行第二运算子步骤s243，1×1运算子模块1213的1×1卷积分配器组可执行1×1参数分配程序s2431以将1×1权重参数分配至不同的1×1本地卷积运算单元1213b中的1×1本地暂存器组1213c以利1×1本地暂存器组1213c执行1×1运算参数产生程序s2432。1×1本地暂存器组1213c可包含二子1×1本地暂存器组1213ca、1213cb。二子1×1本地暂存器组1213ca、1213cb以乒乓的方式作动，进而接收1×1权重参数并输出1×1运算参数至1×1本地滤波运算单元1213d。1×1本地滤波运算单元1213d于1×1卷积运算程序s2433中根据1×1运算参数及3×3后处理运算数据1062进行1×1卷积运算以产生1×1运算数据。于1×1后处理运算程序s2434中，1×1后处理运算单元1213e根据偏压分配器1212于偏压运算子步骤s242中所输出的1×1偏压权重参数及1×1运算数据执行1×1后处理运算以产生1×1后处理运算数据1064。在图5至图7及图12的实施方式中，1×1后处理运算数据1064可为卷积神经网络处理器100的输出数据108。

[0111] 请配合参照图5至图10及图12。详细来说，卷积神经网络处理器100可执行卷积神经网络处理器的数据处理方法s200，且卷积神经网络处理器100包含信息解码单元110及卷积判断单元120。信息解码单元110可执行接收步骤s210、指令解码步骤s220及平行处理步骤s230。解码模块111于接收步骤s210中接收输入程序102后，通过程序存储器1111存储输入程序102，即程序存储子步骤s221，再通过指令解码器1112于程序解码子步骤s222中将存储于程序存储器1111中的输入程序102解码以输出运行指令至平行处理模块112及卷积判断单元120的控制器122，其中输入程序102可包含对应多个输入数据106的相关指令。简单来说，于程序解码子步骤s222中，指令解码器1112将对应其中一输入数据106的相关指令进行解码，以输出运行指令。控制器122于接收运行指令后可根据运行指令控制运算模块121。平行处理模块112于接收步骤s210中接收输入权重参数104，并执行平行处理步骤s230。输入权重参数104包含第一输入权重参数、第二输入权重参数及偏压输入权重参数，第一输入权重参数的数量可为9216的倍数，第二输入权重参数的数量可为1024的倍数，偏压输入权重参数的数量可为64的倍数。换句话说，输入权重参数104包含对应多个输入数据106的相关参数。于权重参数存储子步骤s231中，第一平行子存储器1121aa、第二平行子存储器1121ac及偏压平行子存储器1121ab分别存储第一输入权重参数、第二输入权重参数及偏压输入权重参数，其中第一平行子存储器1121aa为9，第二平行子存储器1121ac及偏压平行子存储器1121ab分别为1。此外，平行处理模块112中的第一平行子处理器1121ba的数量为9，第二平行子处理器1121bc及偏压平行子处理器1121bb的数量均为1。于权重参数处理子步骤s232中，第一平行子处理器1121ba及第二平行子处理器1121bc每周期可处理的第一输入权重参数及第二输入权重参数的数量均为4。第一平行子处理器1121ba及第二平行子处理器1121bc分别需使用256个周期处理与所述其中一输入数据106相对应的第一输入权重参数及第二输入权重参数，以分别输出第一输出权重参数及第二输出权重参数，偏压平行子处理器1121bb则使用64个周期处理与所述其中一输入数据106相对应的偏压输入权重参数以输出偏压输出权重参数。因此，卷积神经网络处理器100可通过执行接收步骤s210、指令解码步骤s220及平行处理步骤s230以平行处理输入权重参数104。

[0112] 卷积判断单元120的运算模块121可执行运算步骤s240，与操作模块121包含3×3运算子模块1211、偏压分配器1212及1×1运算子模块1213。偏压分配器1212可执行偏压运算子步骤s242。于偏压运算子步骤s242中，偏压分配器1212接收3×3偏压权重参数及1×1

偏压权重参数，并将3×3偏压权重参数分配至3×3运算子模块1211中的3×3后处理运算单元1211e，以利3×3后处理运算单元1211e执行3×3后处理运算程序s2414，以及将1×1偏压权重参数分配至1×1运算子模块1213中的1×1后处理运算单元1213e，以利1×1后处理运算单元1213e执行1×1后处理运算程序s2434。

[0113] 3×3运算子模块1211可执行第一运算子步骤s241，且3×3运算子模块1211包含多个3×3卷积分配器组、多个3×3本地卷积运算单元1211b、多个3×3后处理运算单元1211e。3×3卷积分配器组与第一平行子处理器1121ba电性连接，且于3×3参数分配程序s2411中，接收并分配3×3权重参数至3×3本地卷积运算单元1211b，以利3×3本地卷积运算单元1211b执行3×3运算参数产生程序s2412及3×3卷积运算程序s2413。各3×3本地卷积运算单元1211b包含3×3本地暂存器组1211c及3×3本地滤波运算单元1211d。3×3本地暂存器组1211c可执行3×3运算参数产生程序s2412，3×3本地暂存器组1211c包含二子3×3本地暂存器组1211ca、1211cb，并以乒乓的方式执行3×3运算参数产生程序s2412以输出3×3运算参数至3×3本地滤波运算单元1211d。于3×3卷积运算程序s2413中，3×3本地滤波运算单元1211d根据3×3运算参数及输入数据106进行3×3卷积运算以产生3×3运算数据，其中3×3卷积运算的空间滤波位置可分别与第一平行子处理器1121ba中的一者对应。于3×3后处理运算程序s2414中，3×3后处理运算单元1211e根据3×3运算数据及3×3偏压权重参数执行3×3后处理运算以输出3×3后处理运算数据1062。

[0114] 1×1运算子模块1213可执行第二运算子步骤s243，且1×1运算子模块1213包含至少一1×1卷积分配器组、多个1×1本地卷积运算单元1213b、多个1×1后处理运算单元1213e。1×1卷积分配器组与至少一第二平行子处理器1121bc电性连接，且于1×1参数分配程序s2431中，接收并分配1×1权重参数至1×1本地卷积运算单元1213b，以利1×1本地卷积运算单元1213b执行1×1运算参数产生程序s2432及1×1卷积运算程序s2433。各1×1本地卷积运算单元1213b包含1×1本地暂存器组1213c及1×1本地滤波运算单元1213d。1×1本地暂存器组1213c可执行1×1运算参数产生程序s2432，1×1本地暂存器组1213c包含二子1×1本地暂存器组1213ca、1213cb，并以乒乓的方式执行1×1运算参数产生程序s2432以输出1×1运算参数至1×1本地滤波运算单元1213d。于1×1卷积运算程序s2433中，1×1本地滤波运算单元1213d根据1×1运算参数及于3×3后处理运算程序s2414中所产生的3×3后处理运算数据1062进行1×1卷积运算以产生1×1运算数据，其中1×1卷积运算的空间滤波位置可分别与至少一第二平行子处理器1121bc对应。于1×1后处理运算程序s2434中，1×1后处理运算单元1213e根据1×1运算数据及1×1偏压权重参数执行1×1后处理运算以输出1×1后处理运算数据1064。于1×1后处理运算程序s2434中所输出的1×1后处理运算数据1064即为卷积神经网络处理器100执行卷积神经网络处理器的数据处理方法s200所产生的输出数据108。

[0115] 综上所述，卷积神经网络处理器100可通过执行卷积神经网络处理器的数据处理方法s200执行高度平行运算进而提供高性能且低功耗的运算。

[0116] 虽然本发明已以实施方式公开如上，然其并非用以限定本发明，任何本领域技术人员，在不脱离本发明的构思和范围内，当可作各种的变动与润饰，因此本发明的保护范围当视权利要求所界定者为准。

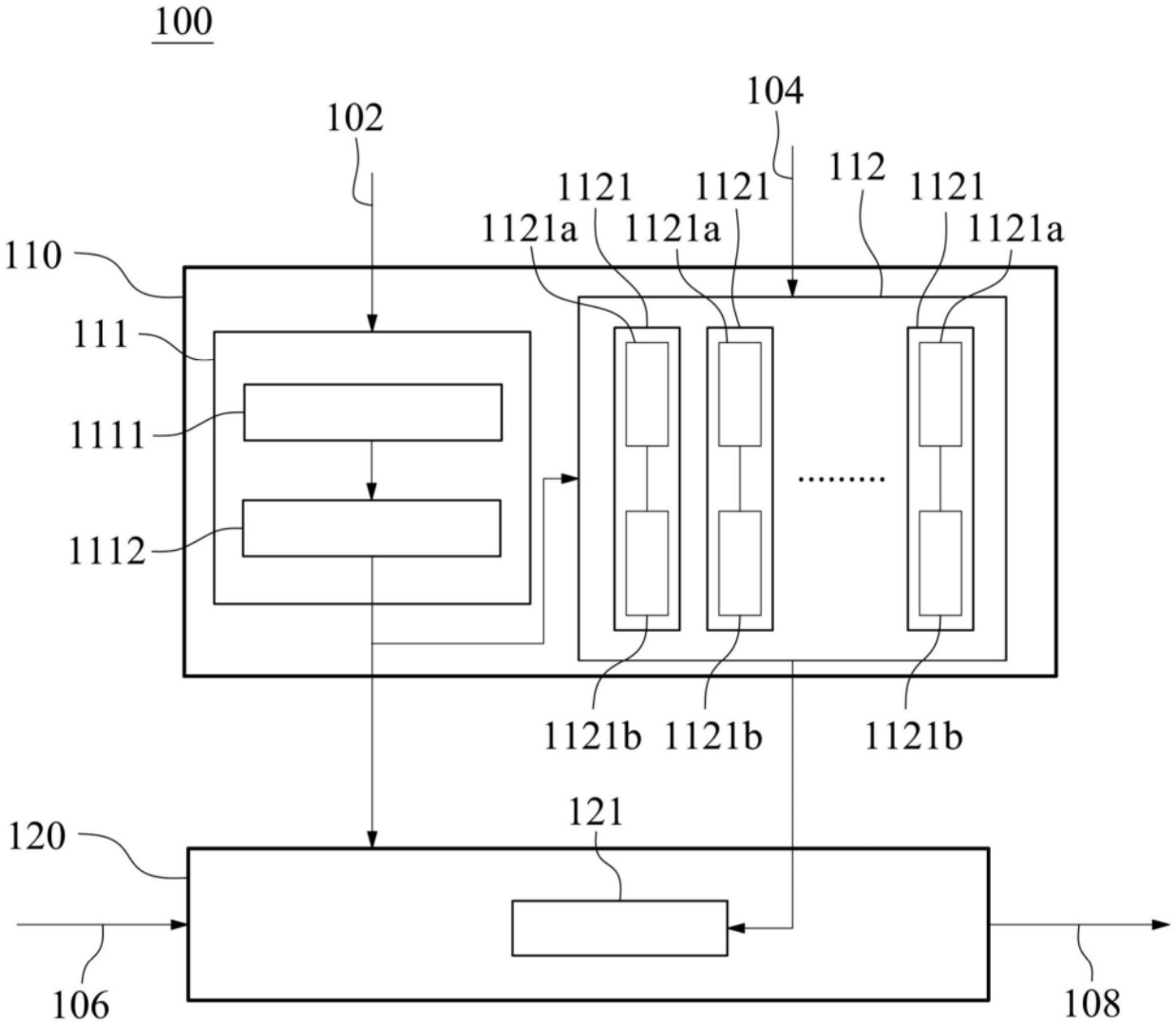

100
102
104
112
1121
1121
1121
1121a
1121a
1121a
110
111
1111
1112
1121b
1121b
1121b
120
121
106
108

图1

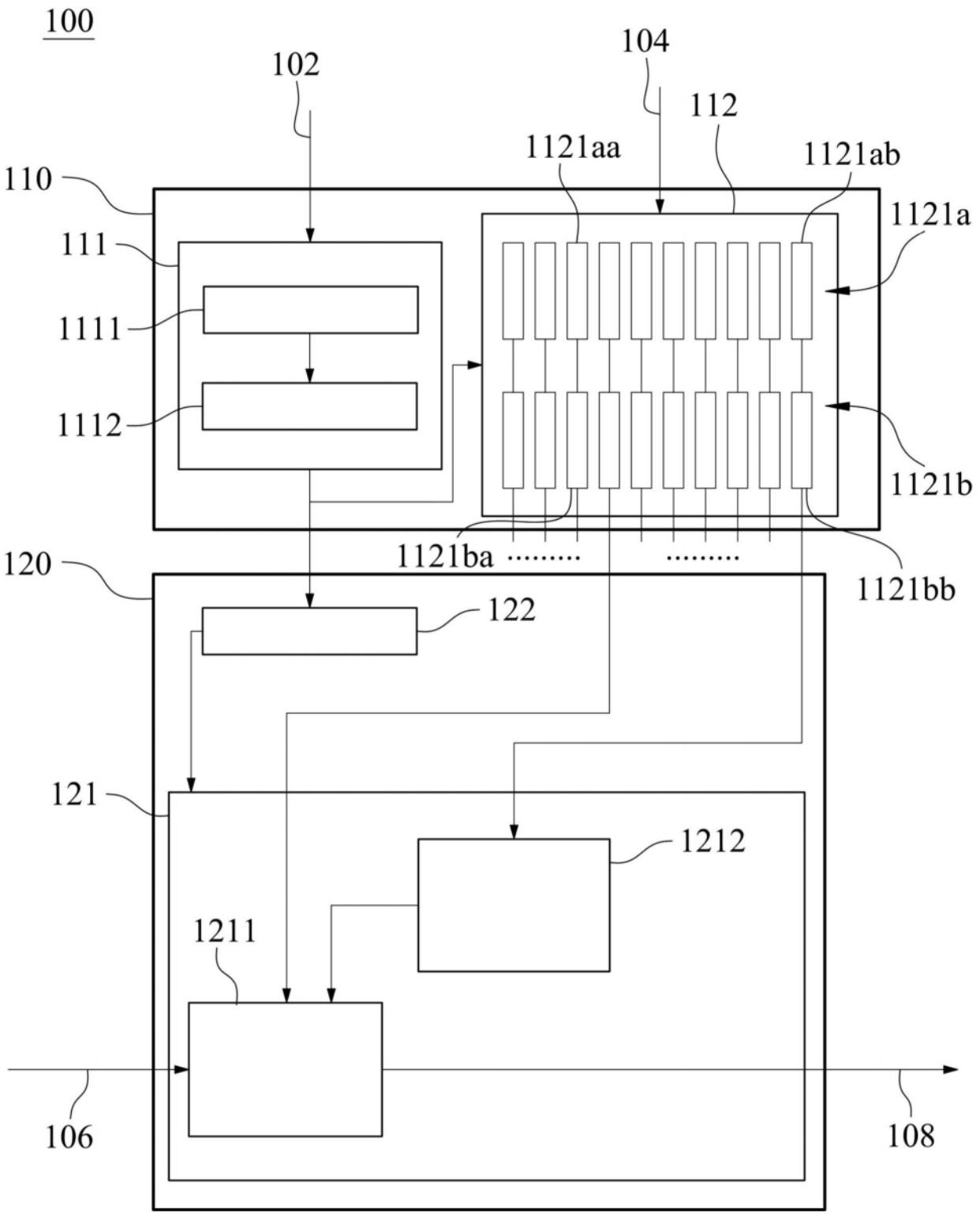

100
102
104
112
1121aa
1121ab
110
1121a
111
1111
1112
1121b
1121ba
1121bb
120
122
121
1212
1211
106
108

图2

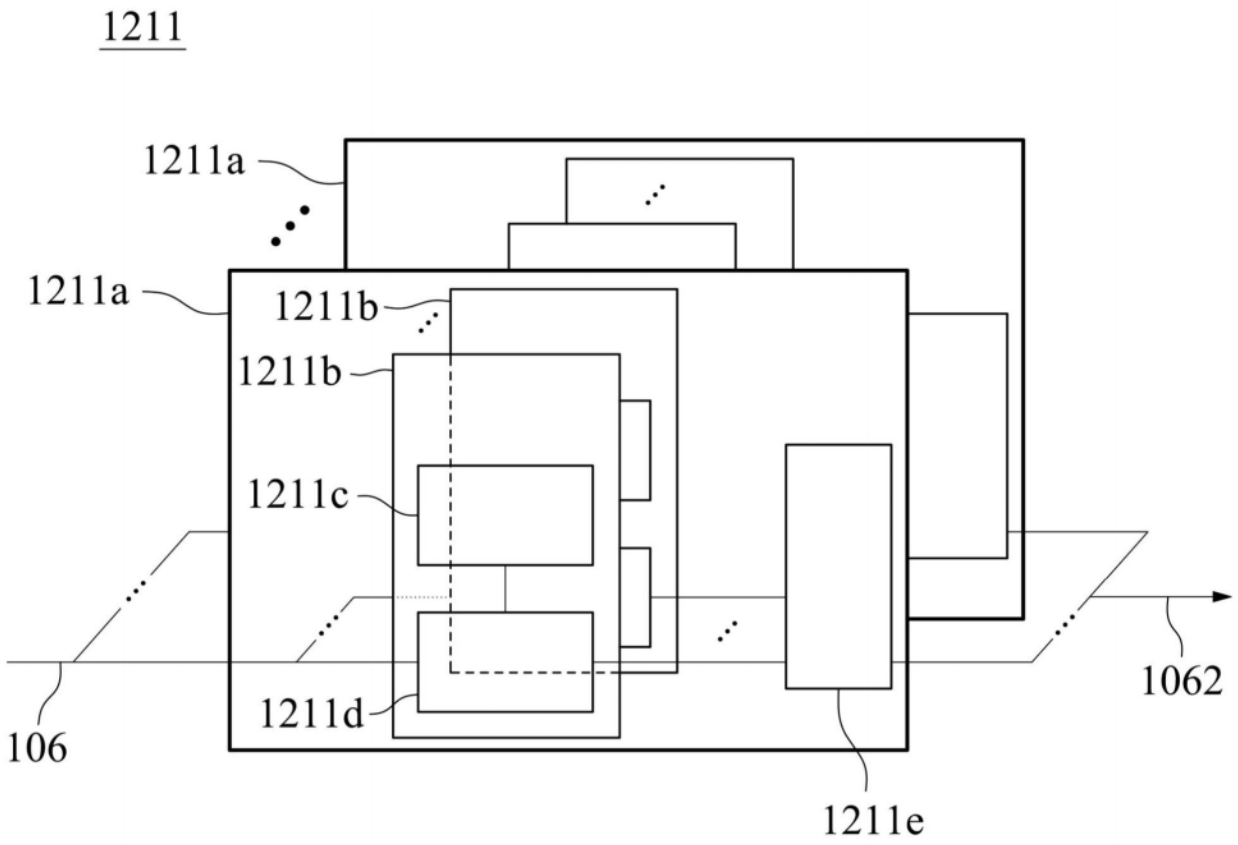


图3

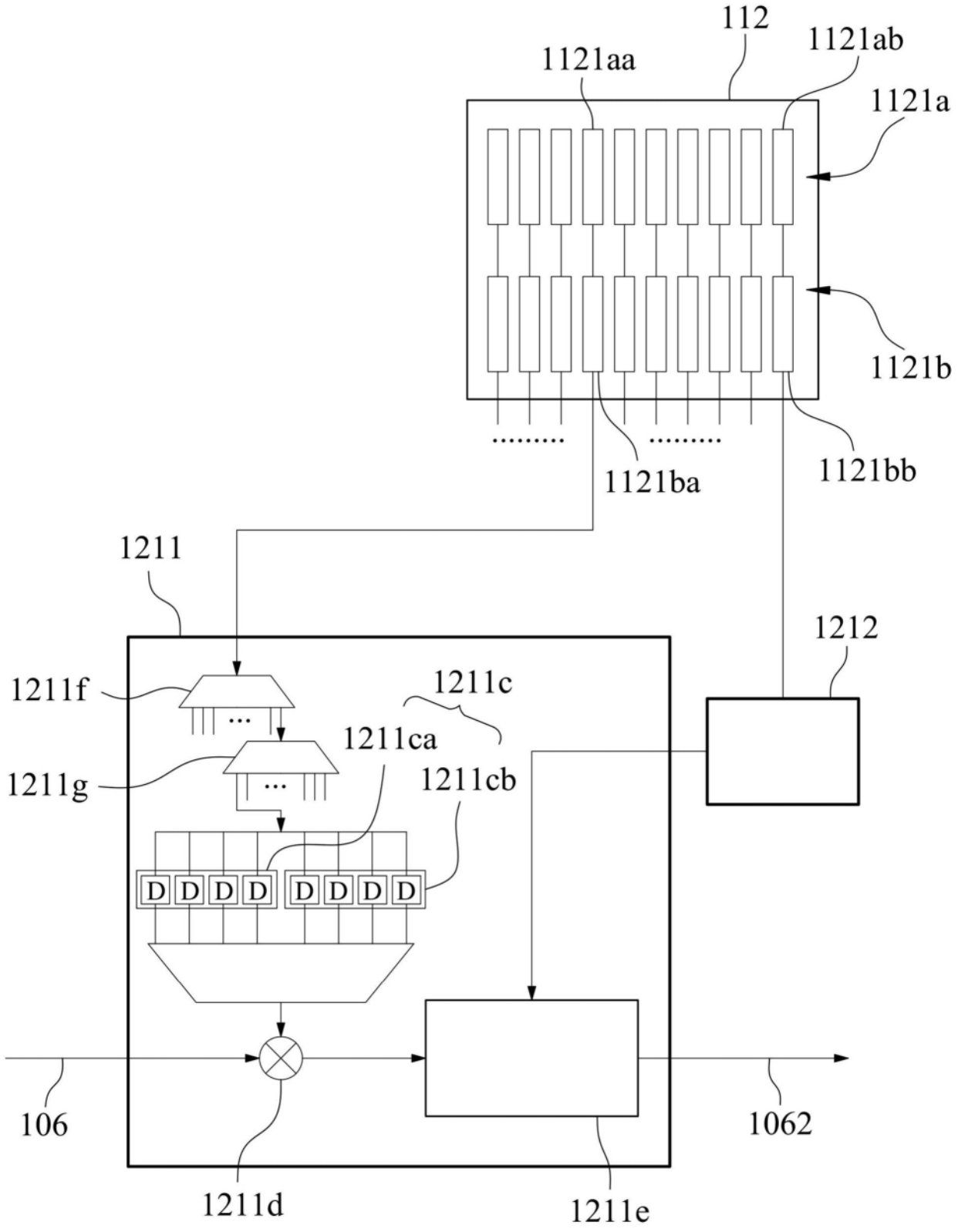

112
1121aa
1121ab
1121a
1121b
1121ba
1121bb
1211
1212
1211f
1211c
1211ca
1211g
1211cb
D D D D D D D D
106
1062
1211d
1211e

图4

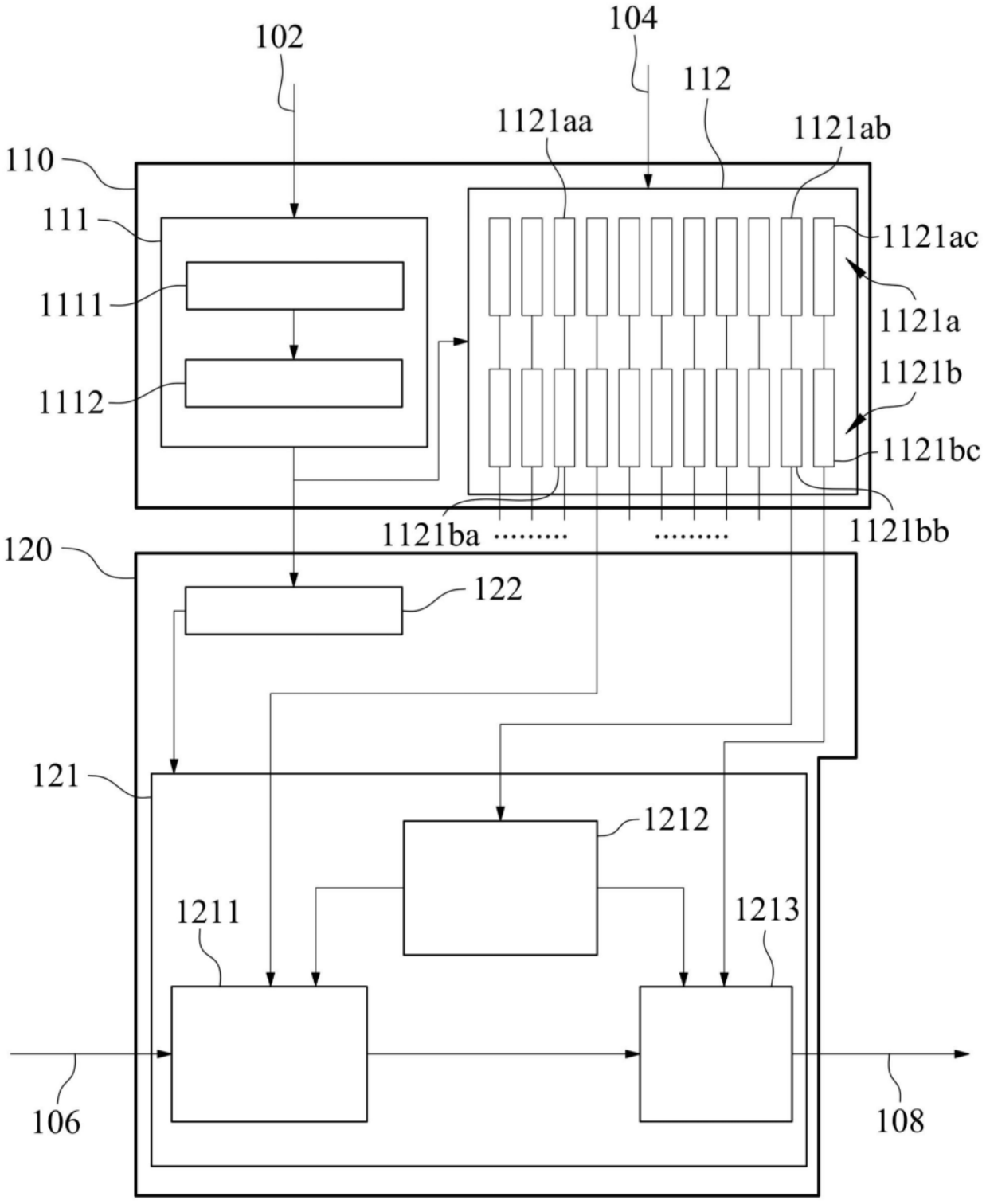

102
104
112
1121aa
1121ab
110
111
1121ac
1111
1121a
1112
1121b
1121bc
1121ba
1121bb
120
122
121
1212
1211
1213
106
108

图5

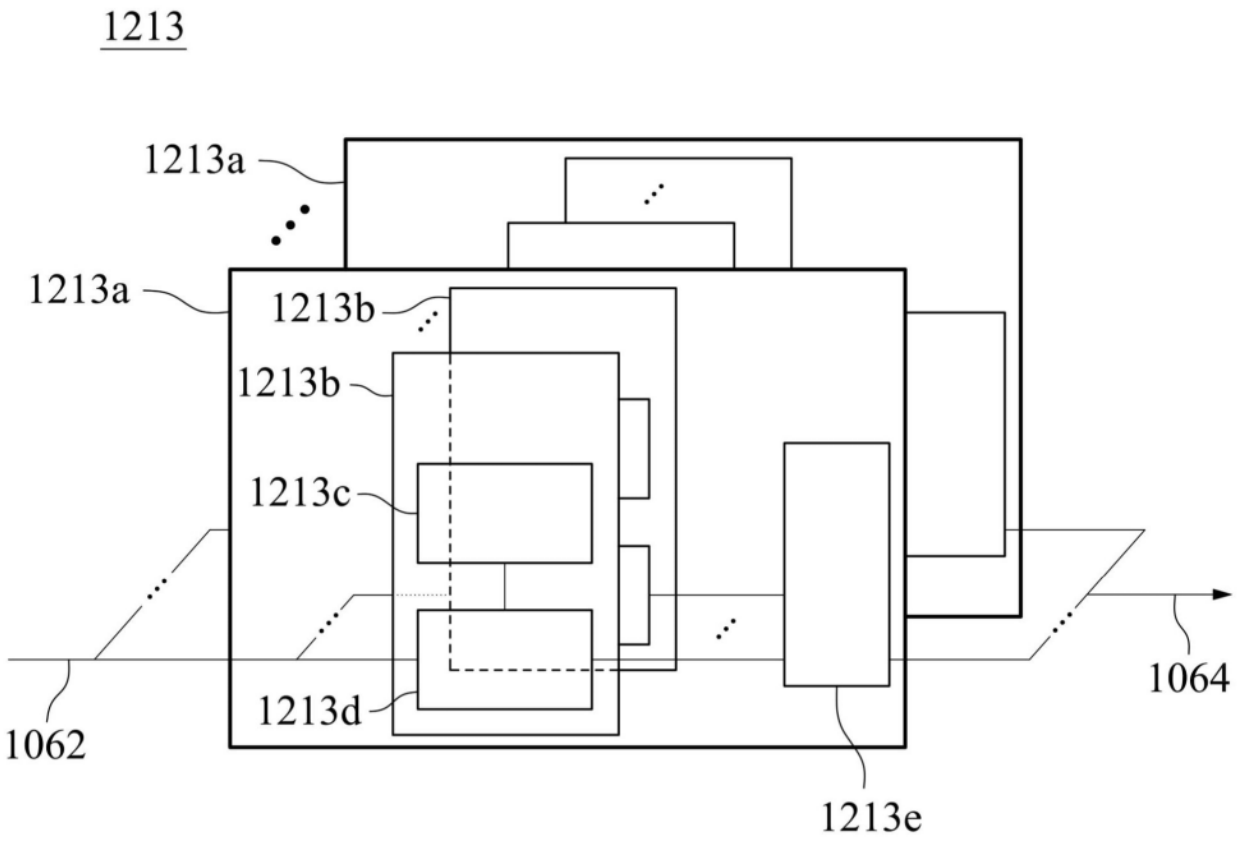


图6

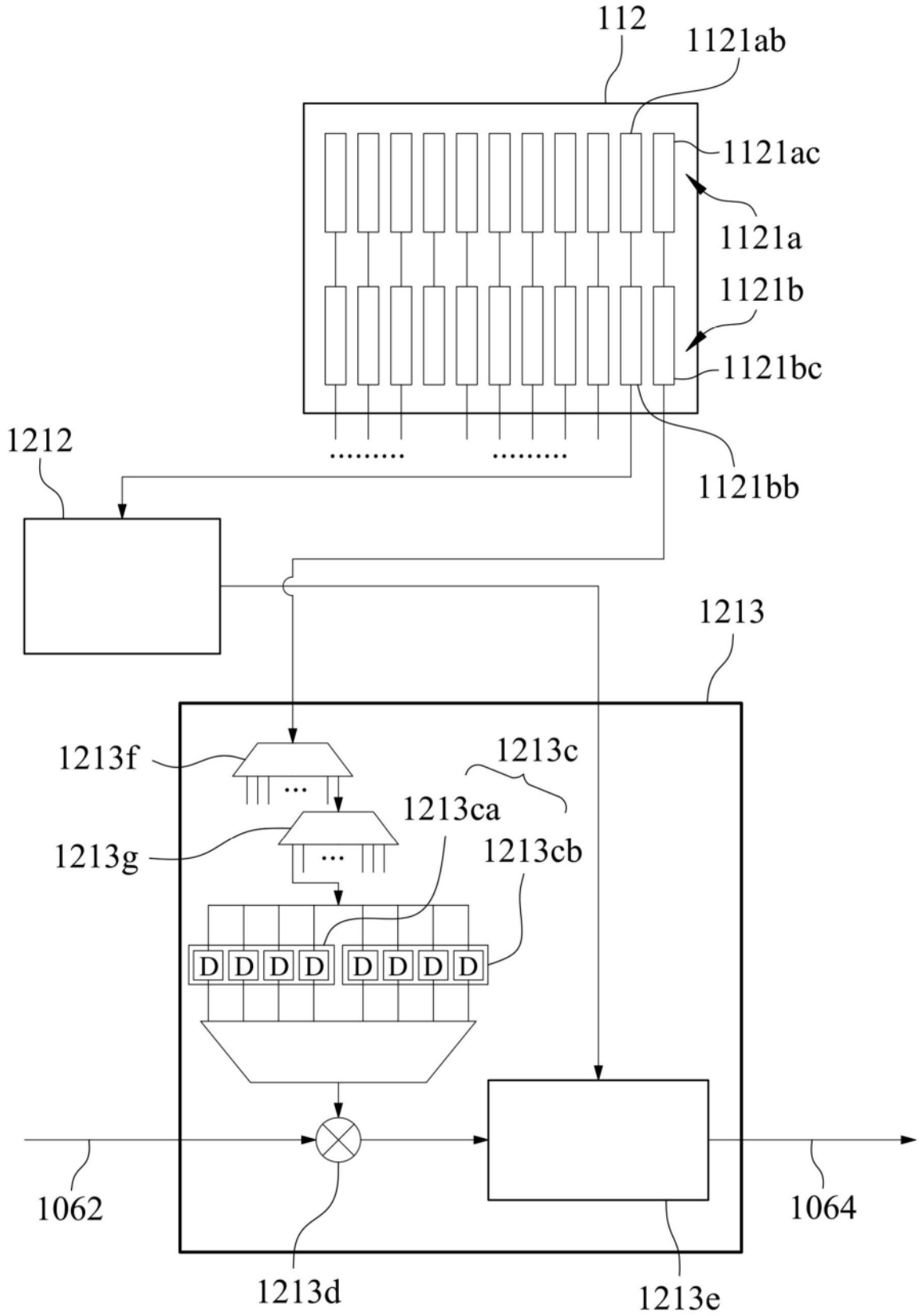

112
1121ab
1121ac
1121a
1121b
1121bc
1121bb
1212
1213
1213f
1213c
1213ca
1213g
1213cb
D D D D D D D D
1062
1064
1213d
1213e

图7

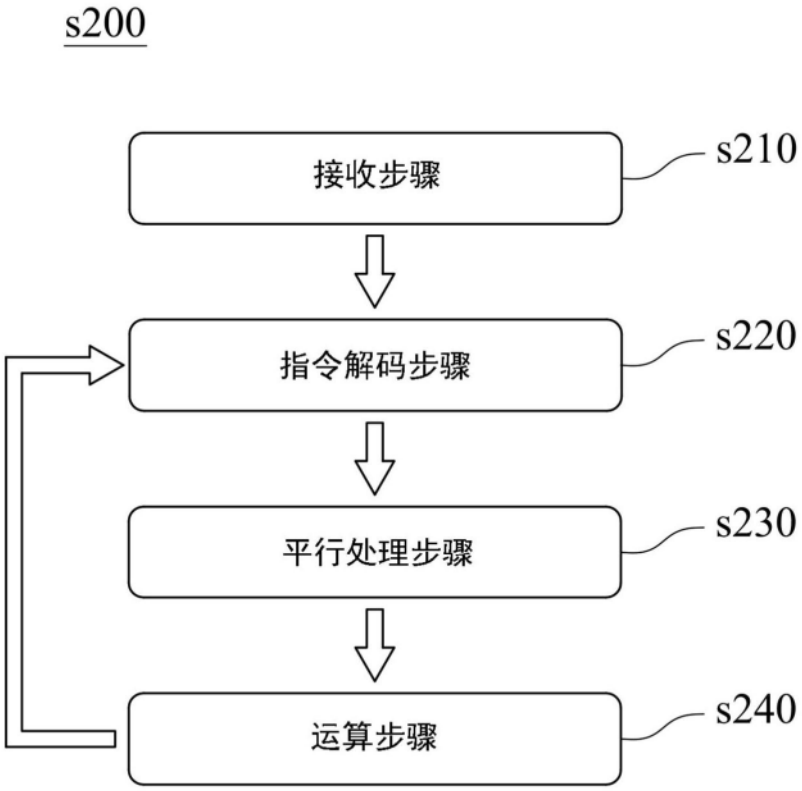

s200
接收步骤
s210
指令解码步骤
s220
平行处理步骤
s230
运算步骤
s240

图8

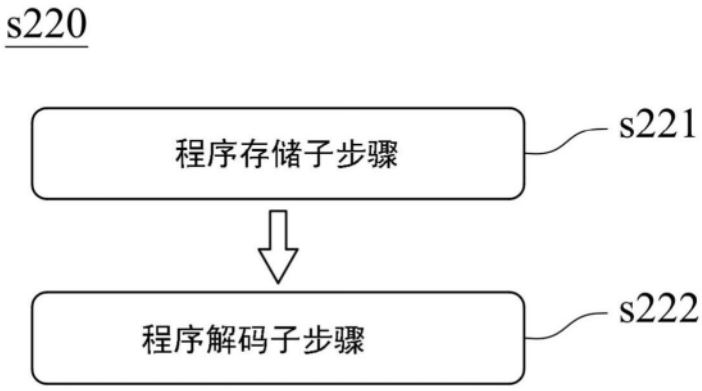

s220
程序存储子步骤
s221
程序解码子步骤
s222

图9

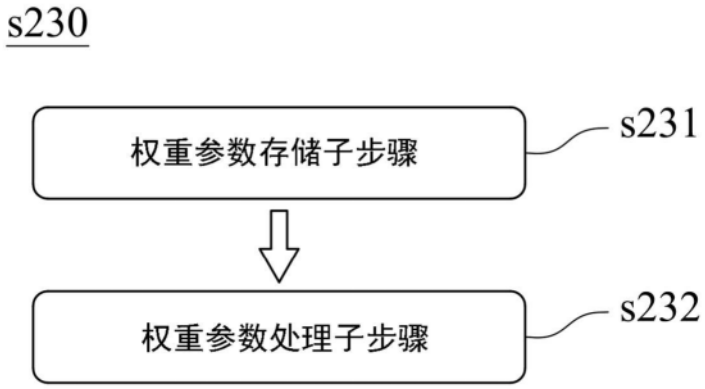

s230
权重参数存储子步骤
s231
权重参数处理子步骤
s232

图10

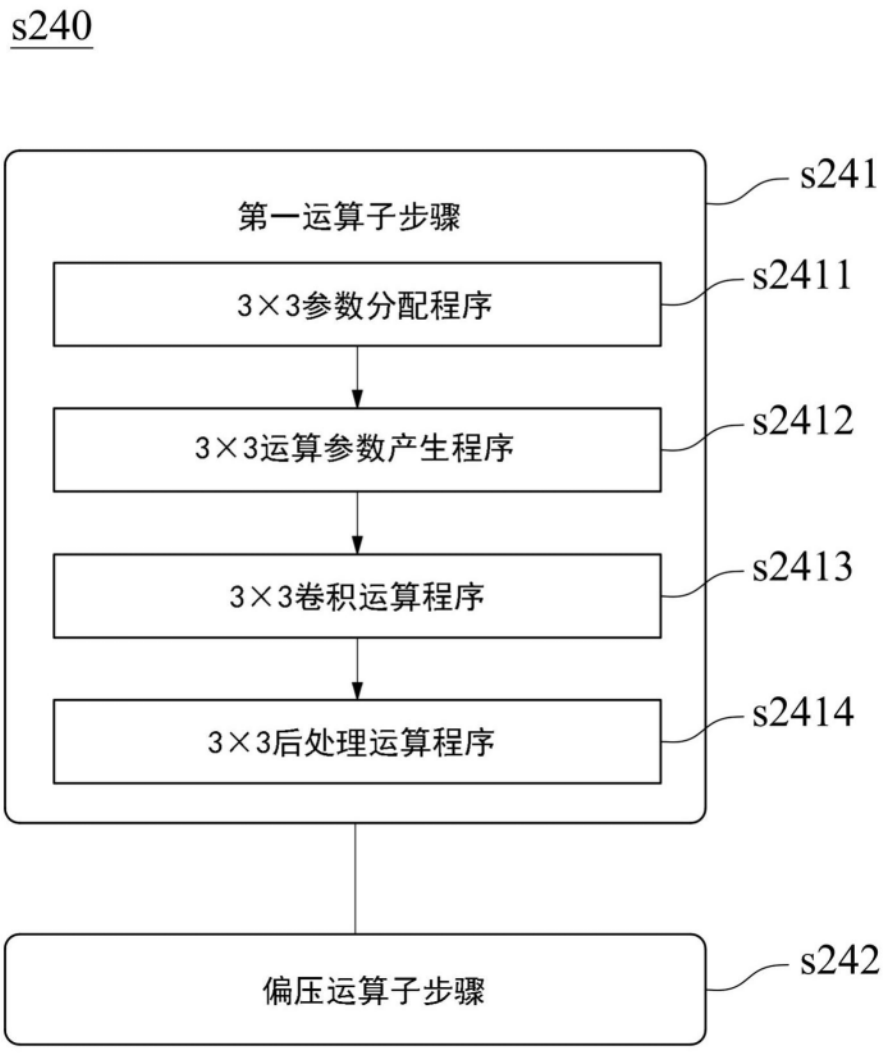

s240
第一运算子步骤
s241
3×3参数分配程序
s2411
3×3运算参数产生程序
s2412
3×3卷积运算程序
s2413
3×3后处理运算程序
s2414
偏压运算子步骤
s242

图11

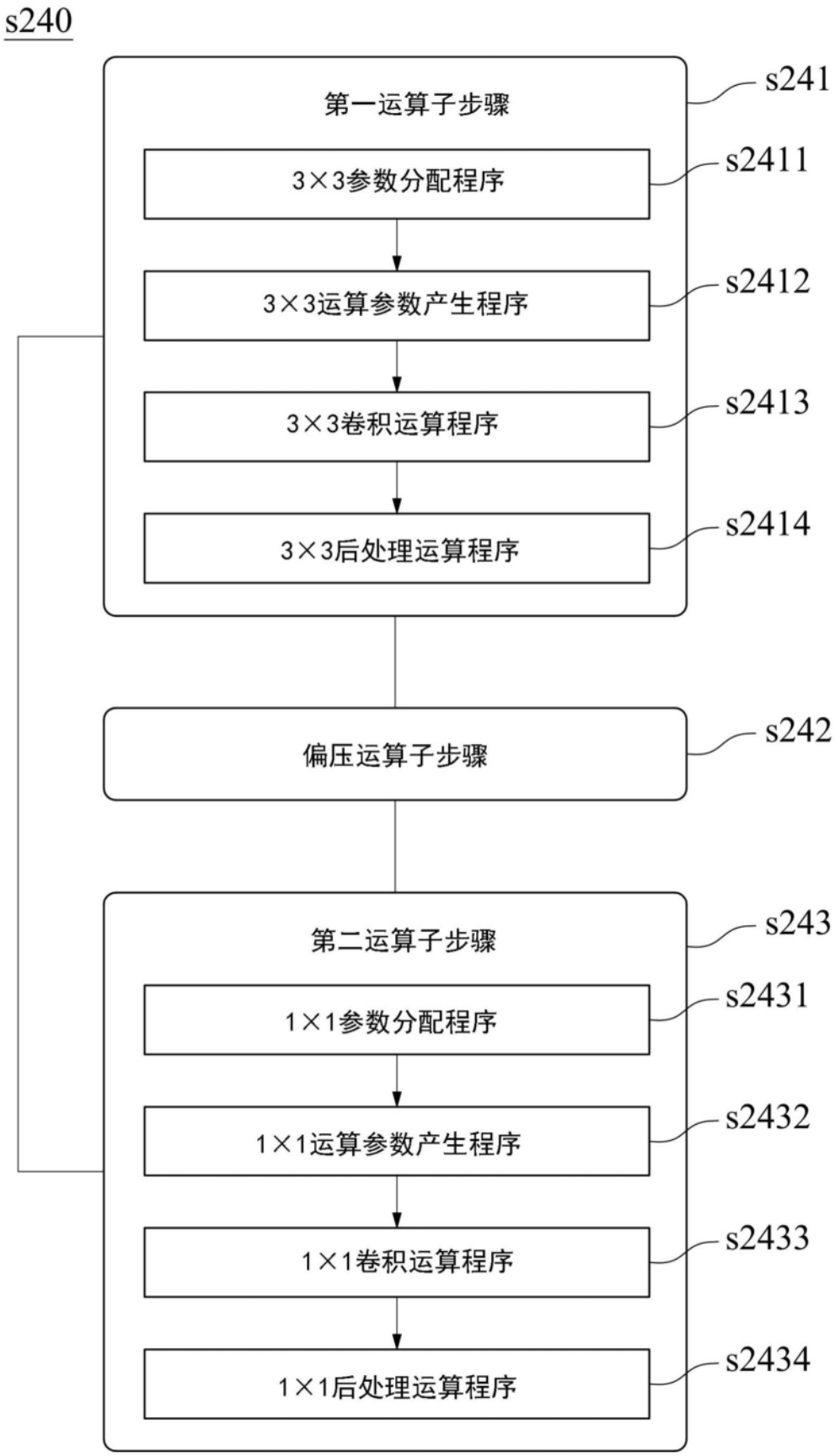

s240
第一运算子步骤
s241
3×3参数分配程序
s2411
3×3运算参数产生程序
s2412
3×3卷积运算程序
s2413
3×3后处理运算程序
s2414
偏压运算子步骤
s242
第二运算子步骤
s243
1×1参数分配程序
s2431
1×1运算参数产生程序
s2432
1×1卷积运算程序
s2433
1×1后处理运算程序
s2434

图12