



(12) 发明专利申请

(10) 申请公布号 CN 113033790 A

(43) 申请公布日 2021.06.25

(21) 申请号 202010599092.5

(22) 申请日 2020.06.28

(30) 优先权数据

10-2019-0162910 2019.12.09 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 蔚山科学技术院

(72) 发明人 俞炯硕 沈贤煜 李宗恩

(74) 专利代理机构 北京铭硕知识产权代理有限公司

公司 11286

代理人 黄晓燕 张川绪

(51) Int.Cl.

G06N 3/063 (2006.01)

G06N 3/04 (2006.01)

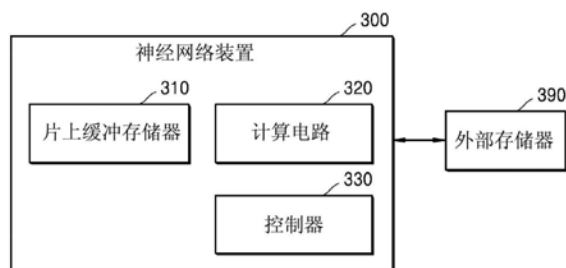
权利要求书3页 说明书17页 附图6页

(54) 发明名称

神经网络装置及其操作方法

(57) 摘要

公开神经网络装置及其操作方法。所述神经网络装置包括：片上缓冲存储器，存储神经网络的第一层的输入特征图；计算电路：通过片上缓冲存储器的单端口接收第一层的输入特征图，并且对第一层的输入特征图执行神经网络运算以输出与第一层的输入特征图对应的第一层的输出特征图；和控制器：通过所述单端口将第一层的输出特征图发送到片上缓冲存储器，以将第一层的输出特征图和第一层的输入特征图一起存储在片上缓冲存储器中。



1. 一种神经网络装置,包括:

片上缓冲存储器,被配置为存储神经网络的第一层的输入特征图;

计算电路,被配置为:通过片上缓冲存储器的单端口接收第一层的输入特征图,并且对第一层的输入特征图执行神经网络运算,以输出与第一层的输入特征图对应的第一层的输出特征图;和

控制器,被配置为:通过所述单端口将第一层的输出特征图发送到片上缓冲存储器,以将第一层的输出特征图和第一层的输入特征图一起存储在片上缓冲存储器中,

其中,第一层的输出特征图被重新用作用于第一层随后的第二层的神经网络运算的输入特征图。

2. 根据权利要求1所述的神经网络装置,其中,计算电路还被配置为:基于一个或多个运算循环来执行神经网络运算,

其中,控制器还被配置为:在每个周期执行通过所述单端口从片上缓冲存储器读取构成第一层的输入特征图的至少部分的数据的读取操作,所述一个或多个运算循环中的每个运算循环在每个周期中执行,并且

当在读取操作将被执行的时间用于通过所述单端口将构成第一层的输出特征图的至少部分的数据写入片上缓冲存储器的写入操作被请求时,优先于读取操作执行写入操作。

3. 根据权利要求1或权利要求2所述的神经网络装置,其中,控制器沿不同的方向分配用于存储第一层的输入特征图的片上缓冲存储器的第一存储器地址和用于存储第一层的输出特征图的片上缓冲存储器的第二存储器地址。

4. 根据权利要求3所述的神经网络装置,其中,控制器:

从与片上缓冲存储器的存储空间对应的存储器地址的起点沿第一方向分配第一存储器地址,并且

从与片上缓冲存储器的存储空间对应的存储器地址的终点沿与第一方向相反的第二方向分配第二存储器地址。

5. 根据权利要求4所述的神经网络装置,其中,

当被存储在第二存储器地址中的第一层的输出特征图被重新用作第二层的输入特征图时,当与第二层的输入特征图对应的第二层的输出特征图从计算电路输出时,控制器从所述起点沿第一方向分配用于存储第二层的输出特征图的片上缓冲存储器的第三存储器地址。

6. 根据权利要求1或权利要求2所述的神经网络装置,其中,神经网络运算包括卷积运算、激活运算和池化运算,

其中,计算电路还被配置为:将对第一层的输入特征图执行池化运算、卷积运算以及激活运算的结果输出为第一层的输出特征图。

7. 根据权利要求1或权利要求2所述的神经网络装置,还包括:

权重缓冲存储器,被配置为:存储用于对第一层的输入特征图的神经网络运算的第一层的权重,

其中,权重缓冲存储器通过权重缓冲存储器的单端口从神经网络装置外部的存储器接收第一层的权重,并且通过权重缓冲存储器的单端口将第一层的权重发送到计算电路。

8. 根据权利要求1所述的神经网络装置, 其中,
片上缓冲存储器、计算电路和控制器安装在单个芯片中。

9. 根据权利要求1或权利要求2所述的神经网络装置, 还包括:
辅助缓冲存储器,

其中, 当与第二层的输入特征图对应的第二层的输出特征图从计算电路输出时, 控制器确定第二层的输入特征图和第二层的输出特征图的总大小是否超过片上缓冲存储器的大小, 并且当所述总大小超过片上缓冲存储器的大小时, 控制器将第二层的输出特征图临时存储在辅助缓冲存储器中而不是片上缓冲存储器中, 并且

其中, 被临时存储在辅助缓冲存储器中的第二层的输出特征图基于预设周期被发送到神经网络装置外部的的外部存储器。

10. 根据权利要求9所述的神经网络装置, 其中,

当第二层的输出特征图被重新用作用于第二层随后的第三层的神经网络运算的输入特征图时, 当与第三层的输入特征图对应的第三层的输出特征图从计算电路输出时, 控制器确定第三层的输出特征图的大小是否超过片上缓冲存储器的大小, 并且

当第三层的输出特征图的大小超过片上缓冲存储器的大小时, 控制器将第三层的输出特征图临时存储在辅助缓冲存储器中, 并且

当第三层的输出特征图的大小小于或等于片上缓冲存储器的大小时, 控制器将第三层的输出特征图存储在片上缓冲存储器中。

11. 一种神经网络装置的操作方法, 所述方法包括:

将神经网络的第一层的输入特征图存储在片上缓冲存储器中;

通过片上缓冲存储器的单端口将第一层的输入特征图发送到计算电路;

在计算电路对第一层的输入特征图执行神经网络运算时, 输出与第一层的输入特征图对应的第一层的输出特征图; 和

通过所述单端口将第一层的输出特征图发送到片上缓冲存储器, 从而将第一层的输出特征图和第一层的输入特征图一起存储在片上缓冲存储器中,

其中, 第一层的输出特征图被重新用作用于第一层随后的第二层的神经网络运算的输入特征图。

12. 根据权利要求11所述的方法, 还包括:

在每个周期执行通过所述单端口从片上缓冲存储器读取构成第一层的输入特征图的至少部分的数据的读取操作, 以基于所述一个或多个运算循环执行神经网络运算, 所述一个或多个运算循环中的每个运算循环在每个周期中执行; 和

当在读取操作将被执行的时间用于通过所述单端口将构成第一层的输出特征图的至少部分的数据写入片上缓冲存储器的写入操作被请求时, 优先于读取操作执行写入操作。

13. 根据权利要求11或权利要求12所述的方法, 其中, 将第一层的输出特征图和第一层的输入特征图一起存储在片上缓冲存储器中的步骤包括:

沿不同的方向分配用于存储第一层的输入特征图的片上缓冲存储器的第一存储器地址和用于存储第一层的输出特征图的片上缓冲存储器的第二存储器地址。

14. 根据权利要求13所述的方法, 其中, 所述分配的步骤包括:

从与片上缓冲存储器的存储空间对应的存储器地址的起点沿第一方向分配用于存储

第一层的输入特征图的第一存储器地址,和

从与片上缓冲存储器的存储空间对应的存储器地址的终点沿与第一方向相反的第二方向分配用于存储第一层的输出特征图的第二存储器地址。

15. 根据权利要求14所述的方法,还包括:

当被存储在第二存储器地址中的第一层的输出特征图被重新用作第二层的输入特征图时,当与第二层的输入特征图对应的第二层的输出特征图从计算电路输出时,从所述起点沿第一方向分配用于存储第二层的输出特征图的片上缓冲存储器的第三存储器地址。

16. 根据权利要求11或权利要求12所述的方法,其中,神经网络运算包括卷积运算、激活运算和池化运算,

其中,所述输出的步骤包括:

将对第一层的输入特征图执行池化运算、卷积运算以及激活运算的结果输出为第一层的输出特征图。

17. 根据权利要求11或权利要求12所述的方法,还包括:

当通过权重缓冲存储器的单端口从神经网络装置外部的的外部存储器将第一层的权重发送到权重缓冲存储器时,将用于神经网络运算的第一层的权重存储在权重缓冲存储器中;和

通过权重缓冲存储器的单端口将第一层的权重从权重缓冲存储器发送到计算电路。

18. 根据权利要求11或权利要求12所述的方法,还包括:

当与第二层的输入特征图对应的第二层的输出特征图从计算电路输出时,确定第二层的输入特征图和第二层的输出特征图的总大小是否超过片上缓冲存储器的大小;和

当确定所述总大小超过片上缓冲存储器的大小时,将第二层的输出特征图临时存储在辅助缓冲存储器中而不是片上缓冲存储器中,

其中,基于预设周期将临时存储在辅助缓冲存储器中的第二层的输出特征图发送到神经网络装置外部的的外部存储器。

19. 根据权利要求18所述的方法,还包括:

当第二层的输出特征图被重新用作第二层随后的第三层的神经网络运算的输入特征图时,当与第三层的输入特征图对应的第三层的输出特征图从计算电路输出时,确定第三层的输出特征图的大小是否超过片上缓冲存储器的大小;和

当第三层的输出特征图的大小超过片上缓冲存储器的大小时,将第三层的输出特征图临时存储在辅助缓冲存储器中,并且当第三层的输出特征图的大小小于或等于片上缓冲存储器的大小时,将第三层的输出特征图存储在片上缓冲存储器中。

20. 一种存储指令的非暂时性计算机可读记录介质,所述指令在被处理器执行时,使处理器控制权利要求11所述的方法的执行。

神经网络装置及其操作方法

[0001] 本申请要求于2019年12月9日在韩国知识产权局提交的第10-2019-0162910号韩国专利申请的权益,所述韩国专利申请的全部公开出于所有目的通过引用包含于此。

技术领域

[0002] 下面的描述涉及神经网络装置和神经网络装置的操作方法。

背景技术

[0003] 神经网络装置是基于计算架构的计算系统。神经网络技术可分析输入数据并从输入数据提取有效信息。

[0004] 神经网络装置通常需要大量的复杂输入数据的计算。为了使典型的神经网络装置处理大量的计算,通常需要执行从存储器读取大量数据或者向存储写入大量数据的操作以用于计算,因此,由于频繁的存储器访问,大量的能量会被消耗。低功率和高性能系统(诸如,移动装置或物联网(IoT)装置)通常具有有限的资源,因此通常需要减少处理大量数据所需的能量消耗的技术。

发明内容

[0005] 提供本发明内容以便以简化的形式介绍以下在具体实施方式中进一步描述的构思的选择。本发明内容不意在确定要求权利的主题的关键特征或必要特征,也不意在用于帮助确定要求权利的主题的范围。

[0006] 在一个总体方面,一种神经网络装置包括片上缓冲存储器、计算电路和控制器。片上缓冲存储器被配置为存储神经网络的第一层的输入特征图。计算电路被配置为:通过片上缓冲存储器的单端口接收第一层的输入特征图,并且对第一层的输入特征图执行神经网络运算,以输出与第一层的输入特征图对应的第一层的输出特征图。控制器被配置为:通过所述单端口将第一层的输出特征图发送到片上缓冲存储器,以将第一层的输出特征图和第一层的输入特征图一起存储在片上缓冲存储器中。第一层的输出特征图被重新用作用于第一层随后的第二层的神经网络运算的输入特征图。

[0007] 计算电路还可被配置为基于一个或多个运算循环来执行神经网络运算。控制器还可被配置为:在每个周期执行通过所述单端口从片上缓冲存储器读取构成第一层的输入特征图的至少部分的数据的读取操作,所述一个或多个运算循环中的每个运算循环在每个周期中被执行。当在读取操作将被执行的时间用于通过所述单端口将构成第一层的输出特征图的至少部分的数据写入片上缓冲存储器的写入操作被请求时,可优先于读取操作执行写入操作。

[0008] 控制器可沿不同的方向分配用于存储第一层的输入特征图的片上缓冲存储器的第一存储器地址和用于存储第一层的输出特征图的片上缓冲存储器的第二存储器地址。

[0009] 控制器可从与片上缓冲存储器的存储空间对应的存储器地址的起点沿第一方向分配第一存储器地址,并且从与片上缓冲存储器的存储空间对应的存储器地址的终点沿与

第一方向相反的第二方向分配第二存储器地址。

[0010] 当被存储在第二存储器地址中的第一层的输出特征图被重新用作第二层的输入特征图时,当与第二层的输入特征图对应的第二层的输出特征图从计算电路输出时,控制器可从所述起点沿第一方向分配用于存储第二层的输出特征图的片上缓冲存储器的第三存储器地址。

[0011] 神经网络运算可包括卷积运算、激活运算和池化运算。计算电路还可被配置为:将对第一层的输入特征图执行池化运算以及卷积运算和激活运算的结果输出为第一层的输出特征图。

[0012] 神经网络装置还可包括:权重缓冲存储器,被配置为:存储用于对第一层的输入特征图的神经网络运算的第一层的权重值。权重缓冲存储器可通过权重缓冲存储器的单端口从神经网络装置外部的的外部存储器接收第一层的权重值,并且通过权重缓冲存储器的单端口将第一层的权重发送到计算电路。

[0013] 片上缓冲存储器、计算电路和控制器可安装在单个芯片中。

[0014] 神经网络装置还可包括辅助缓冲存储器。当与第二层的输入特征图对应的第二层的输出特征图从计算电路输出时,控制器可确定第二层的输入特征图和第二层的输出特征图的总大小是否超过片上缓冲存储器的大小。当所述总大小超过片上缓冲存储器的大小时,控制器可将第二层的输出特征图临时存储在辅助缓冲存储器中而不是片上缓冲存储器中。被临时存储在辅助缓冲存储器中的第二层的输出特征图可基于预设周期被发送到神经网络装置外部的的外部存储器。

[0015] 当第二层的输出特征图被重新用作用于第二层随后的第三层的神经网络运算的输入特征图时,当与第三层的输入特征图对应的第三层的输出特征图从计算电路输出时,控制器可确定第三层的输出特征图的大小是否超过片上缓冲存储器的大小。当第三层的输出特征图的大小超过片上缓冲存储器的大小时,控制器可将第三层的输出特征图临时存储在辅助缓冲存储器中。当第三层的输出特征图的大小小于或等于片上缓冲存储器的大小时,控制器可将第三层的输出特征图存储在片上缓冲存储器中。

[0016] 在另一总体方面,一种神经网络装置的操作方法包括:将神经网络的第一层的输入特征图存储在片上缓冲存储器中;通过片上缓冲存储器的单端口将第一层的输入特征图发送到计算电路;在计算电路对第一层的输入特征图执行神经网络运算时,输出与第一层的输入特征图对应的第一层的输出特征图;和通过所述单端口将第一层的输出特征图发送到片上缓冲存储器,从而将第一层的输出特征图和第一层的输入特征图一起存储在片上缓冲存储器中。第一层的输出特征图被重新用作用于第一层随后的第二层的神经网络运算的输入特征图。

[0017] 所述方法还可包括:在每个周期执行通过所述单端口从片上缓冲存储器读取构成第一层的输入特征图的至少部分的数据的读取操作,以基于所述一个或多个运算循环执行神经网络运算,所述一个或多个运算循环中的每个运算循环在每个周期中执行。当在读取操作将被执行的时间用于通过所述单端口将构成第一层的输出特征图的至少部分的数据写入片上缓冲存储器的写入操作被请求时,优先于读取操作执行写入操作。

[0018] 将第一层的输出特征图和第一层的输入特征图一起存储在片上缓冲存储器中的步骤可包括:沿不同的方向分配用于存储第一层的输入特征图的片上缓冲存储器的第一存

存储器地址和用于存储第一层的输出特征图的片上缓冲存储器的第二存储器地址。

[0019] 所述分配的步骤可包括：从与片上缓冲存储器的存储空间对应的存储器地址的起点沿第一方向分配用于存储第一层的输入特征图的第一存储器地址，和从与片上缓冲存储器的存储空间对应的存储器地址的终点沿与第一方向相反的第二方向分配用于存储第一层的输出特征图的第二存储器地址。

[0020] 当被存储在第二存储器地址中的第一层的输出特征图被重新用作第二层的输入特征图时，当与第二层的输入特征图对应的第二层的输出特征图从计算电路输出时，可从所述起点沿第一方向分配用于存储第二层的输出特征图的片上缓冲存储器的第三存储器地址。

[0021] 神经网络运算可包括卷积运算、激活运算和池化运算。所述输出的步骤可包括：将对第一层的输入特征图执行池化运算以及卷积运算和激活运算的结果输出为第一层的输出特征图。

[0022] 当通过权重缓冲存储器的单端口从神经网络装置外部的的外部存储器将第一层的权重值发送到权重缓冲存储器时，可将用于神经网络运算的第一层的权重值存储在权重缓冲存储器中；并且通过权重缓冲存储器的单端口将第一层的权重值从权重缓冲存储器发送到计算电路。

[0023] 当与第二层的输入特征图对应的第二层的输出特征图从计算电路输出时，可确定第二层的输入特征图和第二层的输出特征图的总大小是否超过片上缓冲存储器的大小。当确定所述总大小超过片上缓冲存储器的大小时，可将第二层的输出特征图临时存储在辅助缓冲存储器中而不是片上缓冲存储器中。可基于预设周期将临时存储在辅助缓冲存储器中的第二层的输出特征图发送到神经网络装置外部的的外部存储器。

[0024] 当第二层的输出特征图被重新用作第二层随后的第三层的神经网络运算的输入特征图时，当与第三层的输入特征图对应的第三层的输出特征图从计算电路输出时，可确定第三层的输出特征图的大小是否超过片上缓冲存储器的大小。当第三层的输出特征图的大小超过片上缓冲存储器的大小时，可将第三层的输出特征图临时存储在辅助缓冲存储器中，并且当第三层的输出特征图的大小小于或等于片上缓冲存储器的大小时，可将第三层的输出特征图存储在片上缓冲存储器中。

[0025] 一种存储指令的非暂时性计算机可读记录介质，所述指令在被处理器执行时，可使处理器控制以上方法的执行。

[0026] 从下面的具体实施方式、附图以及权利要求，其它特征和方面将是清楚的。

附图说明

[0027] 图1是示出神经网络中的输入特征图与输出特征图之间的示例关系的示图。

[0028] 图2是示出示例神经网络架构的示图。

[0029] 图3是示例神经网络装置的框图。

[0030] 图4是示出用于将特征图存储在片上缓冲存储器中的示例存储器地址分配方案的示图。

[0031] 图5是示例神经网络装置的框图。

[0032] 图6是示出神经网络装置的特征图的示例移动路径的示图。

[0033] 图7是示出被执行以执行神经网络运算的示例运算循环的算法。

[0034] 图8是示例电子系统的框图。

[0035] 图9是示出神经网络装置的示例操作方法的流程图。

[0036] 图10是示出神经网络装置的示例操作方法的流程图。

[0037] 贯穿附图和具体实施方式,相同的参考标号表示相同的元件。附图可不按比例,并且为了清楚、说明和方便,附图中的元件的相对尺寸、比例和描绘可被夸大。

具体实施方式

[0038] 提供下面的具体实施方式以帮助读者获得对在此描述的方法、设备和/或系统的全面理解。然而,在理解本申请的公开之后,在此描述的方法、设备和/或系统的各种改变、修改和等同物将是清楚的。例如,在此描述的操作的顺序仅是示例,并且不限于在此阐述的操作的顺序,而是除了必须以特定的顺序发生的操作之外,操作的顺序可如在理解本申请的公开之后将是清楚的那样被改变。此外,为了更加清楚和简明,可省略在理解本申请的公开之后已知的特征的描述。

[0039] 在此描述的特征可以以不同的形式来实现,而不被解释为限于在此描述的示例。相反,在此描述的示例已被提供,以仅示出在理解本申请的公开之后将是清楚的实现在此描述的方法、设备和/或系统的许多可行方式中的一些可行方式。

[0040] 如在此使用的,术语“和/或”包括相关联的所列项中的任何一个以及任何两个或更多的任何组合。

[0041] 尽管在此可使用诸如“第一”、“第二”和“第三”的术语来描述各种构件、组件、区域、层或部分,但是这些构件、组件、区域、层或部分不被这些术语限制。相反,这些术语仅用于将一个构件、组件、区域、层或部分与另一构件、组件、区域、层或部分进行区分。因此,在不脱离示例的教导的情况下,在此描述的示例中所称的第一构件、第一组件、第一区域、第一层或第一部分也可被称为第二构件、第二组件、第二区域、第二层或第二部分。

[0042] 在此使用的术语仅用于描述各种示例,并不用于限制公开。除非上下文另外清楚地指示,否则单数形式也意在包括复数形式。术语“包含”、“包括”和“具有”说明存在陈述的特征、数量、操作、构件、元件和/或它们的组合,但不排除存在或添加一个或多个其它特征、数量、操作、构件、元件和/或它们的组合。

[0043] 如在理解本申请的公开之后将清楚的,在此描述的示例的特征可以以各种方式组合。此外,尽管在此描述的示例具有各种配置,但是如在理解本申请的公开之后将清楚的,其它配置是可行的。

[0044] 图1是示出神经网络中的输入特征图与输出特征图之间的示例关系的示意图。

[0045] 神经网络可以是深度神经网络(DNN)或n层神经网络。DNN或n层神经网络可以是例如卷积神经网络(CNN)、递归神经网络(RNN)、深度信念网络或受限玻尔兹曼机。

[0046] 在图1中,在神经网络的层中,第一特征图FM1可对应于输入特征图,并且第二特征图FM2可对应于输出特征图。特征图可表示表达输入数据的各种特征的数据集。特征图FM1和特征图FM2可具有二维矩阵的元素或三维矩阵的元素,并且像素值可在每个元素中被定义。特征图FM1和特征图FM2中的每个特征图具有宽度W(或被称为列)、高度H(或被称为行)和深度D。深度D可对应于节点的数量。在此,应注意,关于示例或实施例使用术语“可”(例

如,关于示例或实施例可包括或实现什么)表示:存在包括或实现这样的特征的至少一个示例或实施例,而所有示例和实施例不限于此。

[0047] 可对第一特征图FM1和权重图WM执行卷积运算,作为结果,第二特征图FM2可被生成。权重图WM通过使用在权重图WM滤波器的每个元素中定义的权重参数执行与第一特征图FM1的卷积运算来过滤第一特征图FM1的特征。权重图WM被用于在以滑动窗口方式移位第一特征图FM1时,执行与第一特征图FM1的窗口(或被称为区块(tile))的卷积运算。在每个移位期间,包括在权重图WM中的权重参数中的每个权重参数可与第一特征图FM1中的重叠窗口的像素值中的每个像素值相乘并且相乘结果被相加。随着第一特征图FM1和权重图WM被卷积,第二特征图FM2的一个节点可被生成。尽管在图1中仅示出一个权重图WM,但是多个权重图可分别与第一特征图FM1卷积以生成具有多个节点的第二特征图FM2。

[0048] 第二特征图FM2可对应于下一层的输入特征图。例如,第二特征图FM2可以是池化(或子采样)层的输入特征图。

[0049] 图2是示出示例神经网络架构的示意图。

[0050] 在图2中,神经网络2具有包括输入层、隐藏层和输出层的结构。神经网络2可基于接收的输入数据(例如, I_1 和 I_2)来进行运算,并且可基于执行运算的结果来生成输出数据(例如, O_1 和 O_2)。

[0051] 如上所述,神经网络2可以是包括两个或更多个隐藏层的n层神经网络或DNN。例如,如图2中所示,神经网络2可以是包括输入层层1、两个隐藏层层2和层3以及输出层层4的DNN。当神经网络2用DNN架构来实现时,神经网络2包括能够处理有效信息的更多个层,因此可比具有单个层的神经网络处理更高复杂度的数据集。神经网络2被示出为包括四个层,但是这仅是示例,并且神经网络2可包括更少或更多的层,或者可包括更少或更多的节点。也就是说,神经网络2可包括与图2中示出的层不同的各种结构的层。

[0052] 包括在神经网络2中的层中的每个层可包括多个节点。每个节点可对应于多个人工节点(被已知为神经元、PE、单元或类似术语)。例如,如图2中所示,输入层层1可包括两个节点(通道),并且隐藏层层2和层3可均包括三个通道CH1、CH2和CH3。然而,这仅是示例,并且包括在神经网络2中的层中的每个层可包括各种数量的节点(通道)。

[0053] 包括在神经网络2的每个层中的节点可彼此连接以处理数据。例如,一个节点可从其它节点接收数据并对数据进行计算,并且可将计算结果输出到其它节点。

[0054] 每个节点的输入和输出可分别被称为输入激活和输出激活。也就是说,激活可以是一个节点的输出,并且也可以是与包括在下一层中的节点的输入对应的参数。每个节点可基于从包括在前一层中的节点接收的激活、权重和偏置来确定每个节点自己的激活。权重是用于计算每个节点中的输出激活的参数,并且可以通过迭代训练针对节点之间的连接关系被分配的值。

[0055] 每个节点可作为计算单元或处理元件来处理,或者由计算单元或处理元件来处理,计算单元或处理元件接收输入并输出输出激活,所以每个节点的输入-输出可被映射。例如,当 σ 是激活函数、 w_{jk}^i 是从包括在第(i-1)层中的第k个节点到包括在第i层中的第j节点的权重、 b_j^i 是包括在第i层中的第j节点的偏置、并且 a_j^i 是包括在第i层中的第j节点的激活时,激活 a_j^i 可使用如下的等式1来计算。

[0056] 等式1:

$$[0057] \quad a_j^i = \sigma\left(\sum_k (w_{jk}^i \times a_k^{i-1}) + b_j^i\right)$$

[0058] 如图2中所示,第二层(即,隐藏层层2)的第一节点示出的激活可由 a_1^2 表示。根据等式1, a_1^2 可具有 $a_1^2 = \sigma(w_{1,1}^2 \times a_1^1 + w_{1,2}^2 \times a_2^1 + b_1^2)$ 的值。然而,以上描述的等式1仅是用于描述用于在神经网络2中处理数据的激活、权重和偏置的示例,但不限于此。激活可以通过将从前一层接收的激活的加权和传递到激活函数(诸如,S型(sigmoid)函数或修正线性单元(ReLU)函数)而获得的值。

[0059] 如上所述,在神经网络2中,大量数据集在多个互连节点之间被交换并经由多个层进行了多个计算处理。因此,需要一种能够减少多个计算处理所需的读取或写入大量数据所需的功耗的技术。

[0060] 为了描述的方便,图1和图2仅示出神经网络2的示意性架构。然而,与图1和图2中示出的架构不同,神经网络2可用更多或更少的层、特征图、权重图等来实现,并且其大小也在各种其它示例中被各种修改。

[0061] 图3是示例神经网络装置300的框图。

[0062] 在图3中,神经网络装置300可包括片上缓冲存储器310、计算电路320和控制器330,并且神经网络装置300可向外部存储器390发送数据并从外部存储器390接收数据,外部存储器390在神经网络装置300的外部。神经网络装置300是具有片上结构的装置,并且神经网络装置300中的组件可安装在一个芯片中。

[0063] 神经网络装置300可以是设计成适于执行神经网络的硬件加速器。神经网络装置300可用于提高包括神经网络装置300作为加速器的电子系统的处理速度。

[0064] 在图3中示出的神经网络装置300中,除了图3中示出的组件之外,神经网络装置300还可包括其它通用组件。

[0065] 片上缓冲存储器310表示设置在与神经网络装置300对应的芯片中的存储器。(例如,与典型的神经网络装置相比)片上缓冲存储器310可通过将输入特征图和输出特征图一起存储来减少特征图到神经网络装置300的移动。因此,在这样的典型的示例中,读取/写入输入特征图或输出特征图所需的对外部存储器390的访问可被减少。

[0066] 片上缓冲存储器310可被配置为一个存储器地址空间。存储器地址空间是定义用于存储数据的存储器地址的范围的空间,并且存储器地址空间表示可存储分配有存储器地址的特征图的空间。例如,当与片上缓冲存储器310的存储器地址空间对应的存储器地址的范围是0x0000至0xFFFF时,输入特征图或输出特征图被分配与0x0000至0xFFFF的至少部分对应的存储器地址,因此可被存储在片上缓冲存储器310中。典型的神经网络装置中设置的典型的存储器包括多个存储器地址空间。存储输入特征图的存储器地址空间与存储输出特征图的存储器地址空间分离。相比之下,根据一个或多个实施例的片上缓冲存储器310可将输入特征图和输出特征图一起存储在一个存储器地址空间中。

[0067] 与典型的神经网络装置相比,片上缓冲存储器310可将输入特征图和输出特征图一起存储在一个存储器地址空间中,因此可减小片上缓冲存储器310的容量,以减小神经网络装置300占用的面积。在将输入特征图和输出特征图存储在分开的缓冲存储器中或者存

储在一个缓冲存储器内的分开的存储器地址空间中的典型的存储器中,需要能够容纳输入特征图的最大大小和输出特征图的最大大小之和的空间。

[0068] 例如,当第一层的输入特征图和输出特征图的大小分别为2兆字节(MB)和6MB、并且第二层的输入特征图和输出特征图的大小分别为7MB和1MB时,典型的缓冲存储器必须具有能够容纳至少13MB数据的容量。相比之下,作为一个非限制性示例,在上述示例中,根据实施例的片上缓冲存储器310可仅具有能够容纳8MB数据的容量。将进一步参照图4描述用于将输入特征图和输出特征图一起存储在片上缓冲存储器310的一个地址空间中的存储器地址分配方法。

[0069] 片上缓冲存储器310使用单端口。片上缓冲存储器310对单端口的使用可表示用于对特征图进行读取操作的端口和用于对特征图进行写入操作的端口是相同的。片上缓冲存储器310可通过连接到一个存储器地址空间的单端口来移动存储在一个存储器地址空间中的特征图。与具有相同容量的使用双端口的典型的缓冲存储器相比,使用单端口的片上缓冲存储器310可具有一半的面积和功耗。换句话说,当使用单端口的片上缓冲存储器310的面积和使用双端口的典型的缓冲存储器的面积相同时,使用单端口的片上缓冲存储器310可具有使用双端口的典型的缓冲存储器的容量的两倍的容量。

[0070] 片上缓冲存储器310的单端口的宽度可根据特征图的存储单元来确定。单端口的宽度可由TN字(TN words)来确定。TN是特征图的并行度(degree of parallelism)。并行度是表示构成特征图并可在单个运算中被处理的特征的单位的变量。字是指表示与一个特征对应的数据所需的位数。例如,与一个字对应的位数可根据数据的小数点格式来确定。小数点格式可在神经网络装置300的各种示例中被选择为浮点格式或定点格式。

[0071] 在一个实施例中,片上缓冲存储器310可存储包括在神经网络中的第一层的输入特征图。第一层不限于神经网络中的第一层,而是表示包括在神经网络中的若干层中的任何一层,并且是用于将一个层与其它层区分开的术语。

[0072] 计算电路320可以是用于通过对输入特征图执行神经网络运算来输出与输入特征图对应的输出特征图的硬件配置。例如,计算电路320可通过片上缓冲存储器310的单端口接收第一层的输入特征图,并且对输入特征图执行神经网络运算以输出第一层的输出特征图。

[0073] 在一个示例中,神经网络运算可包括卷积运算、激活运算和池化运算(pooling operation)。计算电路320可针对对输入特征图进行卷积运算的结果执行激活运算,对激活运算的结果执行池化运算,并将池化运算的结果作为输出特征图输出。在一个示例中,计算电路320中的所有三种运算可在一个芯片中执行,而无需访问外部存储器。除了以上描述的运算之外,计算电路320的运算可包括用于神经网络的运算的各种运算(诸如,批量归一化(batch normalization))。

[0074] 控制器330可起到控制神经网络装置300的总体作用。例如,控制器330可控制片上缓冲存储器310的操作。控制器330可被实现为多个逻辑门的阵列,或者可被实现为通用微处理器和存储可在微处理器中执行的指令的存储器的组合。

[0075] 控制器330可通过片上缓冲存储器310的单端口将存储在片上缓冲存储器310中的输入特征图从片上缓冲存储器310发送到计算电路320。控制器330可通过片上缓冲存储器310的单端口将输出特征图发送到片上缓冲存储器310,输出特征图是对输入特征图进行神

神经网络运算的输出结果。

[0076] 控制器330可将输入特征图和输出特征图一起存储在片上缓冲存储器310中。例如,控制器330可通过片上缓冲存储器310的单端口将第一层的输出特征图发送到片上缓冲存储器310,并且将第一层的输出特征图和第一层的输入特征图一起存储在片上缓冲存储器310中。第一层的输出特征图可被重新用作用于第二层的神经网络运算的输入特征图,第二层是第一层的下一层。

[0077] 例如,与典型的神经网络装置相比,当使用片上缓冲存储器310重新使用特征图时,由于特征图仅在具有片上结构的神经网络装置300中被移动和存储,所以对外部存储器390的用于特征图的写入和读取的访问可被排除。

[0078] 图4是示出用于将特征图存储在片上缓冲存储器中的示例存储器地址分配方案的示意图。

[0079] 为了将输入特征图和输出特征图一起存储在一个存储器地址空间中,神经网络装置(例如,图3的神经网络装置300)采用调整用于向片上缓冲存储器分配特征图的存储器地址的点和方向的方法。

[0080] 神经网络装置可以以相反的方向分配用于存储特定层中的输入特征图和输出特征图的存储器地址。例如,当输入特征图和输出特征图中的任何一个特征图的存储器地址从存储器地址的起点(存储器地址:0)开始被分配并沿朝向存储器地址的终点(存储器地址:最大)的方向被分配时,输入特征图和输出特征图中的另一个特征图的存储器地址可从存储器地址的终点开始被分配并沿朝向存储器地址的起点的方向被分配。

[0081] 在图4的示例中,用于存储第一层的输入特征图(IFM)的片上缓冲存储器的第一存储器地址410可从存储器地址的起点开始被分配并沿朝向存储器地址的终点的方向被分配,并且用于存储第一层的输出特征图(OFM)的片上缓冲存储器的第二存储器地址420可从存储器地址的终点开始被分配并沿朝向存储器地址的起点的方向被分配。

[0082] 当存储在第二存储器地址420中的第一层的输出特征图被重新用作第二层的输入特征图时,与第二层的输入特征图对应的第二层的输出特征图可从计算电路输出。用于存储第二层的输出特征图的片上缓冲存储器的第三存储器地址430可从存储器地址的起点开始被分配并沿朝向存储器地址的终点的方向被分配。

[0083] 由于根据图4的存储器地址分配方案,神经网络装置可将输入特征图和输出特征图一起存储,同时减小在一个存储器地址空间中的输入特征图与输出特征图之间的重叠。神经网络装置可通过将输入特征图和输出特征图两者存储在片上缓冲存储器中来减少对外部存储器的访问。特征图的存储和移动两者可在具有片上结构的神经网络装置中被执行。

[0084] 如上所述,神经网络装置可在具有片上结构的神经网络装置内部移动并存储特征图,从而减少访问外部存储器所产生的功耗。然而,当基于输入特征图和输出特征图的总大小难以将输入特征图和输出特征图一起存储在片上缓冲存储器中时,神经网络装置可在兼容模式下操作。在下文中,将参照图5进一步描述不管图4的存储器地址分配方案,而是可不将输入特征图和输出特征图一起存储在片上缓冲存储器中的情况。

[0085] 图5是示例神经网络装置500的框图。

[0086] 在图5中,神经网络装置500可包括片上缓冲存储器510、计算电路520、控制器530

以及辅助缓冲存储器550。神经网络装置500可向外部存储器590发送数据并从外部存储器590接收数据,外部存储器590在神经网络装置500的外部。神经网络装置500是具有片上结构的装置,并且神经网络装置500中的组件可安装在一个芯片中。图5的片上缓冲存储器510、计算电路520和控制器530可分别对应于图3的片上缓冲存储器310、计算电路320和控制器330。因此,省略重复的描述。

[0087] 控制器530可确定输入特征图和输出特征图的总大小是否超过片上缓冲存储器510的大小。当由于输入特征图和输出特征图的总大小超过片上缓冲存储器510的大小,神经网络装置500可不将输入特征图和输出特征图一起存储在片上缓冲存储器510中时,神经网络装置500可选择在兼容模式下操作。在兼容模式下,神经网络装置500可使用除片上缓冲存储器510之外的辅助缓冲存储器550来存储特征图。

[0088] 辅助缓冲存储器550可选择仅在兼容模式下操作。当从计算电路520输出的输出特征图可能未被存储在片上缓冲存储器510中时,辅助缓冲存储器550可临时容纳输出特征图。辅助缓冲存储器550可临时容纳输出特征图,并且临时容纳在辅助缓冲存储器550中的输出特征图可根据预设周期被发送到外部存储器590。输出特征图可被存储在外部存储器590中,可从外部存储器590传送到计算电路520,并且可被重新用于下一层的神经网络运算的输入特征图。

[0089] 图6是示出神经网络装置的特征图的示例移动路径的示意图。

[0090] 在图6中,神经网络装置包括权重缓冲存储器640、片上缓冲存储器610、辅助缓冲存储器650和局部总线660。神经网络装置还可包括例如计算电路和控制器。图6的片上缓冲存储器610、计算电路、辅助缓冲存储器650以及控制器可分别对应于图5的片上缓冲存储器510、计算电路520、辅助缓冲存储器550以及控制器530,因此,省略其重复的描述。计算电路可包括卷积运算电路621、激活运算电路622和池化运算电路623。外部存储器690在神经网络装置的外部。

[0091] T_z 是输入特征图的并行度,并且是表示构成输入特征图并可在单个运算中处理的特征的单位的变量。 T_m 是输出特征图的并行度,并且是表示构成输出特征图并可在单个运算中处理的特征的单位的变量。

[0092] 在向神经网络装置的第一层(输入层)输入外部数据的处理、从神经网络装置的最后一层(输出层)输出数据的处理、以及读取权重的处理中,神经网络装置对外部存储器690的访问被进行。在其它层中,神经网络装置的操作(诸如,移动、计算和存储特征图)可全部在具有片上结构的神经网络装置内部执行,而无需访问外部存储器690。

[0093] 权重缓冲存储器640可存储用于对输入特征图进行神经网络运算的权重。权重缓冲存储器640可使用单端口。权重缓冲存储器640可通过单端口从外部存储器690接收权重。权重缓冲存储器640可通过单端口将权重发送到卷积运算电路621,以对输入特征图进行神经网络运算。卷积运算电路621可基于权重对输入特征图执行卷积运算。

[0094] 可将存储在片上缓冲存储器610中的输入特征图发送到卷积运算电路621,并且可基于权重对输入特征图执行卷积运算。可将输入特征图执行卷积运算的结果输入到激活运算电路622,然后可执行激活运算。可将执行激活操作的结果输入到池化运算电路623,并且可直接执行池化运算。作为执行池化运算的结果而输出的输出特征图可被发送到片上缓冲存储器610,并且与输入特征图一起被存储。然而,当输入特征图和输出特征图的总大小

超过片上缓冲存储器610的大小,并且因此神经网络装置可不将输入特征图和输出特征图一起存储在片上缓冲存储器610中时,神经网络装置可在兼容模式下操作。当神经网络装置选择在兼容模式下操作时,辅助缓冲存储器650可进行操作。当存储在片上缓冲存储器610中的输入特征图和从池化运算电路623输出的输出特征图的总大小超过片上缓冲存储器610的大小时,控制器(例如,图3的控制器330或图5的控制器530)可将输出特征图临时容纳在辅助缓冲存储器650中。输出特征图可被临时容纳在辅助缓冲存储器650中,并且可根据预设周期被发送到外部存储器690。

[0095] 例如,当存储在片上缓冲存储器610中的第二层的输入特征图和从计算电路输出的第二层的输出特征图的总大小超过片上缓冲存储器610的大小时,控制器可将第二层的输出特征图临时容纳在辅助缓冲存储器650中。

[0096] 当输入特征图被存储在外部存储器690中时,控制器可基于输出特征图的大小是否超过片上缓冲存储器610的大小来确定输出特征图的存储位置。例如,当第二层的输出特征图被存储在外部存储器690中并且第二层的输出特征图被重新用作第三层的输入特征图时,控制器可确定从计算电路输出的第三层的输出特征图的大小是否超过片上缓冲存储器610的大小。当第三层的输出特征图的大小超过片上缓冲存储器610的大小时,控制器可将第三层的输出特征图临时容纳在辅助缓冲存储器650中。当第三层的输出特征图的大小小于或等于片上缓冲存储器610的大小时,控制器可将第三层的输出特征图存储在片上缓冲存储器610中。

[0097] 然而,即使当神经网络装置选择在兼容模式下操作并将特征图存储在外部存储器690中时,对于特征图的再使用所需的特征图的最小部分也可被存储在片上缓冲存储器610中,并且剩余部分可被存储在外部存储器690中。

[0098] 神经网络装置可通过在计算电路内处理池化运算而不是在单独的层中处理池化运算来减小输出特征图的大小。例如,可通过池化运算将输出特征图的大小减小池化步长的平方。因此,将输出特征图存储在片上缓冲存储器610中的可能性可被增大,并且对外部存储器690的访问可被减少。

[0099] 局部总线660是缓冲存储器与使用单端口的其它组件之间的数据(特征图或权重)的移动路径。特征图可经由局部总线660在片上缓冲存储器610与计算电路之间移动。权重可经由局部总线660在权重缓冲存储器640与外部存储器690或计算电路之间移动。当大量数据必须经由单端口来发送时,局部总线660可防止单端口上的冲突(collision)。当特征图的读取操作和特征图的写入操作被同时请求时,读取操作和写入操作的顺序可在局部总线660上调整,因此,读取操作与写入操作之间的冲突可在单端口中被防止。

[0100] 图7是示出被执行以执行神经网络运算的示例运算循环的算法。

[0101] M表示输出特征图中的特征的数量,R和C表示输出特征图中的行的数量和列的数量,Z表示输入特征图中的特征的数量,K表示卷积滤波器的大小,S表示卷积的步长,P表示池化的步长,Q表示池化滤波器的大小。

[0102] 神经网络装置可执行至少一个运算循环以执行神经网络运算。例如,神经网络装置可用输入特征图作为输入来执行M-loop。M-loop是指定将被输出的输出特征图的特征的数量的范围的循环。特征的数量的范围可在每个循环中以 T_M 个元素为单位进行扩展。M-loop可包括R/C-loop。R/C-loop是指定在输出输出特征图的处理中将被输出的输出特征图

的行和列的循环。R/C-loop包括RR-loop、CC-loop、R-loop以及C-loop。

[0103] 因为R/C-loop在M-loop中,所以权重缓冲存储器的大小可被确定为相对小。R/C-loop在M-loop中表示:当以将被输出的输出特征图中的特征的数量固定范围扩展行和列以及输出输出特征图的特征的处理被完成时,在将被输出的特征的数量范围移动 T_M 个元素的同时扩展行和列以及输出输出特征图的特征的处理被重复。

[0104] 在一个示例中,使运算循环内的权重的重新使用最大化所需的权重缓冲存储器的大小可与 $T_M \times Z \times K^2$ 成比例。与根据图7的算法不同,该大小小于当M-loop处于R/C-loop中时权重缓冲存储器所需的大小 $M \times Z \times K^2$ 。由于根据图7的算法,神经网络装置中权重缓冲存储器所占用的面积可被减小。

[0105] K-loop#1是对输入特征图和权重执行卷积运算的循环。紧在K-loop#1的卷积运算被执行之后,激活运算和池化运算被连续执行。由于在一个循环中连续处理激活运算和池化运算而不是在单独的循环中处理激活运算和池化运算,因此可简化神经网络运算的处理。当R/C-loop在没有其它循环的干预的情况下连续进行时,池化运算可容易地在循环中处理。

[0106] 每当在卷积运算中使用的输入特征图的特征的范围被改变时,可作出对输入特征图的读取请求。也就是说,在输入特征图的z值在Z-loop中被改变的每个循环中,可请求读取操作。每当行或列被改变时,可做出对输出特征图的写入请求。也就是说,每当R/C-loop中的r或c值改变时(当K-loop#1的循环完成时),可请求写入操作。可在每个周期作出对其z值改变的输入特征图的读取请求,并且可在每 $K \times K \times Z/T_Z$ 周期作出对其r值或c值改变的输出特征图的写入请求。

[0107] 当读取操作每个周期被执行并且写入操作每 $K \times K \times Z/T_Z$ 个周期被执行时,每 $K \times K \times Z/T_Z$ 个周期写入操作与读取操作之间的冲突可在单端口处发生。然而,与每个周期发生的读取请求的频率相比,每 $K \times K \times Z/T_Z$ 个周期发生的写入请求的频率可对应于非常低的频率。例如,以 $K \times K \times Z/T_Z$ 个周期在单端口处发生的冲突的频率在整个循环进行重复的频率的1%以内,并且对神经网络装置的运算速度的影响可能非常小。

[0108] 行700和行710是仅当神经网络装置选择在兼容模式下操作时使用的代码。行700是用于当输入特征图被存储在外部存储器中时从外部存储器读取输入特征图的代码。行710是用于将输出特征图写入外部存储器的代码。

[0109] 图8是示例电子系统800的框图。电子系统800可对应于参照图1至图7以及图9和图10描述的任何神经网络装置或所有神经网络装置。

[0110] 在图8中,电子系统800可通过基于神经网络实时分析输入数据来提取有效信息,并且可基于提取的信息进行情况确定或控制安装有电子系统800的电子装置的配置。例如,电子系统800可应用于或者是机器人装置(诸如,无人机和高级驾驶员辅助系统(ADAS))、智能电视、智能电话、医疗装置、移动装置、图像显示装置、测量装置,物联网(IoT)装置等。另外,电子系统800可安装在各种类型的电子装置中的至少一个电子装置上。

[0111] 电子系统800可包括中央处理器(CPU) 810、随机存取存储器(RAM) 820、神经网络装置830、存储器840、传感器模块850以及通信模块860。电子系统800还可包括输入/输出模块、安全模块、电源控制装置等。电子系统800的硬件配置中的一些可以是至少一个半导体芯片或者安装在至少一个半导体芯片上。神经网络装置830可以是具有以上描述的片上结

构的任何神经网络装置或所有神经网络装置 (例如,图3的神经网络装置300或图5的神经网络装置500) 或包括神经网络装置的设备。

[0112] CPU 810控制电子系统800的整体操作。CPU 810可包括一个处理器核或多个处理器核。CPU 810可处理或执行存储在存储器840中的指令和/或数据。在一个实施例中,CPU 810可通过执行存储在存储器840中的程序来控制神经网络装置830的功能。CPU 810可被实现为CPU、图形处理器 (GPU)、应用处理器 (AP) 等。

[0113] RAM 820可临时存储程序、数据或指令。例如,存储在存储器840中的指令和/或数据可在CPU 810或启动代码的控制下临时存储在RAM 820中。RAM 820可被实现为存储器 (诸如,动态随机存取存储器 (DRAM) 或静态随机存取存储器 (SRAM))。

[0114] 神经网络装置830可基于接收的输入数据来执行神经网络的运算,并且基于执行的结果来生成信息信号。神经网络可包括但不限于CNN、RNN、深度信念网络、受限玻尔兹曼机等。神经网络装置830可对应于专用于神经网络的硬件加速器。

[0115] 信息信号可包括各种类型的识别信号 (诸如,语音识别信号、对象识别信号、图像识别信号以及生物特征信息识别信号) 之一。例如,神经网络装置830可接收包括在视频流中的帧数据作为输入数据,并且从帧数据生成针对包括在由帧数据表示的图像中的对象的识别信号。然而,本公开不限于此,神经网络装置830可根据安装有电子系统800的电子装置的类型或功能接收各种类型的输入数据,并可根据输入数据来生成识别信号。

[0116] 存储器840是用于存储数据的存储位置,并且可存储操作系统 (OS)、各种指令和各种类型的数据。存储器840可对应于图3的外部存储器390。当针对每个操作将在神经网络装置830的运算期间生成的中间结果 (例如,输出特征图) 存储在存储器840中时,由于对存储器840的频繁访问,功耗可增大。神经网络装置830可通过将在神经网络装置830的运算期间生成的特征图的存储限制到神经网络装置830,来减少对存储器840的访问。作为结果,根据公开的神经网络装置830可减少功耗。存储器840还可存储在神经网络装置830中使用的量化的神经网络数据 (诸如,参数、权重图或权重列表)。

[0117] 存储器840可以是DRAM,但不限于此。存储器840可包括易失性存储器和非易失性存储器中的至少一个。非易失性存储器包括只读存储器 (ROM)、可编程ROM (PROM)、可擦除可编程ROM (EPROM)、电可擦除和可编程ROM (EEPROM)、闪存、相变RAM (PRAM)、磁RAM (MRAM)、电阻式RAM (RRAM)、铁电RAM (FRAM) 等。易失性存储器包括DRAM、SRAM、同步DRAM (SDRAM)、PRAM、MRAM、RRAM、FRAM等。在一个实施例中,存储器840可包括硬盘驱动器 (HDD)、固态驱动器 (SSD)、紧凑型闪存 (CF)、安全数字 (SD)、微型SD、迷你SD、极限数字 (xD) 和记忆棒中的至少一个。

[0118] 传感器模块850可收集安装有电子系统800的电子装置外围的信息。传感器模块850可从电子装置的外部感测或接收信号 (例如,图像信号、音频信号、磁信号、生物信号或触摸信号),并且将感测或接收的信号转换为数据。为此,传感器模块850可包括各种类型的感测装置 (诸如,麦克风、成像装置、图像传感器、光检测和测距 (LIDAR) 传感器、超声传感器、红外传感器、生物传感器以及触摸传感器) 中的至少一种。

[0119] 传感器模块850可将转换的数据作为输入数据提供给神经网络装置830。例如,传感器模块850可包括图像传感器,并且可通过拍摄电子装置的外部环境来生成视频流,并且按顺序将视频流的连续数据帧作为输入数据提供给神经网络装置830。然而,本公开不限于

此,并且传感器模块850可向神经网络装置830提供各种类型的数据。

[0120] 通信模块860可包括用于与外部装置通信的各种有线接口或无线接口。例如,通信模块860可包括可连接到有线局域网(LAN)、无线局域网(WLAN)(诸如,无线保真(Wi-Fi))、无线个域网(WPAN)(诸如,蓝牙)、无线通用串行总线(USB)、zigbee、近场通信(NFC)、射频识别(RFID)、电力线通信(PLC)或移动蜂窝网络(诸如,第三代(3G)、第四代(4G)或长期演进(LTE))的通信接口。

[0121] 图9是示出神经网络装置的示例操作方法的流程图。

[0122] 在图9中,神经网络装置的操作方法包括在图3中示出的神经网络装置300或图5中示出的神经网络装置500中以时间顺序处理的操作。因此,即使以下被省略,以上参照图3至图8给出的神经网络装置的描述也可应用于图9的方法。

[0123] 在操作910中,神经网络装置可将神经网络中的第一层的输入特征图存储在片上缓冲存储器中。

[0124] 在操作920中,神经网络装置可通过片上缓冲存储器的单端口将第一层的输入特征图发送到计算电路。

[0125] 在操作930中,神经网络装置可通过对第一层的输入特征图执行神经网络运算来输出与第一层的输入特征图对应的第一层的输出特征图。

[0126] 神经网络装置可基于一个或多个运算循环来执行神经网络运算。神经网络装置可在每个周期执行通过单端口从片上缓冲存储器读取构成第一层的输入特征图的至少部分的数据的读取操作,在每个周期中一个或多个运算循环中的每个运算循环被执行。然而,当在读取操作将被执行的时间用于通过单端口将构成第一层的输出特征图的至少部分的数据写入片上缓冲存储器的写入操作被请求时,可优先于读取操作执行写入操作。

[0127] 神经网络运算可包括卷积运算、激活运算和池化运算,神经网络装置可将第一层的输入特征图执行池化运算、卷积运算以及激活运算的结果输出为第一层的输出特征图。

[0128] 在操作940中,神经网络装置可通过单端口将第一层的输出特征图发送到片上缓冲存储器,并且将第一层的输出特征图和第一层的输入特征图一起存储在片上缓冲存储器中。

[0129] 神经网络装置可沿不同的方向分配用于存储第一层的输入特征图的片上缓冲存储器的第一存储器地址和用于存储第一层的输出特征图的片上缓冲存储器的第二存储器地址,从而减小第一存储器地址与第二存储器地址之间的重叠。

[0130] 神经网络装置可从与片上缓冲存储器的存储空间对应的存储器地址的起点沿第一方向分配第一存储器地址,并且从与片上缓冲存储器的存储空间对应的存储器地址的终点沿与第一方向相反的第二方向分配第二存储器地址。

[0131] 当存储在第二存储器地址中的第一层的输出特征图被重新用作第二层的输入特征图时,当与第二层的输入特征图对应的第二层的输出特征图从计算电路输出时,神经网络装置可从起点沿第一方向分配用于存储第二层的输出特征图的片上缓冲存储器的第三存储器地址,从而减小第二存储器地址与第三存储器地址之间的重叠。

[0132] 在一个实施例中,神经网络装置可将用于对第一层的输入特征图的神经网络运算的第一层的权重存储在权重缓冲存储器中。权重缓冲存储器可通过权重缓冲存储器的单端

口从神经网络装置外部的的外部存储器接收第一层的权重。权重缓冲存储器可通过权重缓冲存储器的单端口将第一层的权重发送到计算电路。

[0133] 片上缓冲存储器、计算电路和神经网络装置的控制器可安装在单个芯片中。图9中的每个操作可在单个芯片中执行。

[0134] 在一个示例中,当与第二层的输入特征图对应的第二层的输出特征图从计算电路输出时,神经网络装置可确定第二层的输入特征图和第二层的输出特征图的总大小是否超过片上缓冲存储器的大小。当确定第二层的输入特征图和第二层的输出特征图的总大小超过片上缓冲存储器的大小时,神经网络装置可将第二层的输出特征图临时存储在辅助缓冲存储器而不是片上缓冲存储器中。可根据预设周期将被临时存储在辅助缓冲存储器中的第二层的输出特征图发送到神经网络装置外部的的外部存储器。

[0135] 当第二层的输出特征图被重新用作用于第三层的神经网络运算的输入特征图时,神经网络装置可从计算电路输出与第三层的输入特征图对应的第三层的输出特征图,第三层是第二层的下一层。

[0136] 神经网络装置可确定第三层的输出特征图的大小是否超过片上缓冲存储器的大小。当确定第三层的输出特征图的大小超过片上缓冲存储器的大小时,神经网络装置可将第三层的输出特征图临时存储在辅助缓冲存储器中。当确定第三层的输出特征图的大小小于或等于片上缓冲存储器的大小时,神经网络装置可将第三层的输出特征图存储在片上缓冲存储器中。

[0137] 图10是示出神经网络装置的示例操作方法的流程图。

[0138] 在图10中,神经网络装置的操作方法包括在图3中示出的神经网络装置300或图5中示出的神经网络装置500中以时间顺序处理的操作。因此,即使以下被省略,以上参照图3至图8给出的神经网络装置的描述也可应用于图10的方法。

[0139] 根据公开的神经网络装置的原理操作方法是输入特征图和输出特征图一起存储在片上缓冲存储器中的方法。然而,输入特征图和输出特征图可不被一起存储在片上缓冲存储器中的例外情况可能发生。为了对这样的例外情况作出准备,除了原理操作方法之外,根据公开的神经网络装置可在兼容模式下操作。当输入特征图和输出特征图的总大小超过片上缓冲存储器的大小时,作为例外,神经网络装置可在兼容模式下操作。图10是示出包括神经网络装置的原理操作方法和兼容模式的整体操作处理的示图。

[0140] 图10中的操作S1031、S1071、S1072以及S1073表示神经网络装置选择在兼容模式下操作的情况下的操作。

[0141] 在图10中,示出从神经网络装置的第一层(输入层)到神经网络装置的最后层(输出层)的操作方法。图10中表示的‘n’指示神经网络装置中的层的顺序。例如,当神经网络装置包括总共五个层时,‘n=1’表示第一层,‘n=2’表示第二层,并且‘n=5’表示最后层。

[0142] 在操作1010和操作1020中,神经网络装置可将当外部数据被输入时首先生成的第一层(n=1)的输入特征图存储在外部存储器中。当在操作1030中确定第一层的输入特征图的大小小于或等于片上缓冲存储器的大小时,神经网络装置可执行操作1040。在操作1040中,神经网络装置可将第一层的输入特征图存储在片上缓冲存储器中。当在操作1030中确定第一层的输入特征图的大小超过片上缓冲存储器的大小时,神经网络装置可执行操作S1031。在操作S1031中,神经网络装置可在兼容模式下操作以将第一层的输入特征图存储

在外部存储器中。在操作1050中,神经网络装置可将第一层的输入特征图发送到计算电路以执行神经网络运算并输出第一层的输出特征图。

[0143] 当在操作1060中确定第一层是最后层时,神经网络装置可执行操作1120。在操作1120中,神经网络装置可将第一层的输出特征图存储在外部存储器中。当在操作1060中确定第一层不是最后层时,神经网络装置可执行操作1070。当在操作1070中确定第一层的输入特征图和输出特征图的总大小小于或等于片上缓冲存储器的大小时,神经网络装置可执行操作1080。在操作1080中,神经网络装置可将第一层的输出特征图存储在片上缓冲存储器中。当在操作1070中确定第一层的输入特征图和输出特征图的总大小超过片上缓冲存储器的大小时,神经网络装置可进入操作S1071并在兼容模式下操作。

[0144] 当在操作S1071中确定第一层的输入特征图根据操作1040被存储在片上缓冲存储器中时,神经网络装置可执行操作S1072。在操作S1072中,神经网络装置可将第一层的输出特征图临时存储在辅助缓冲存储器中。在操作S1072中,可将临时存储在辅助缓冲存储器中的第一层的输出特征图传送到外部存储器并存储在外部存储器中。当在操作S1071中确定第一层的输入特征图根据操作S1031被存储在外部存储器中时,神经网络装置可进入操作S1073。

[0145] 当在操作S1073中确定第一层的输出特征图的大小小于或等于片上缓冲存储器的大小时,神经网络装置可执行操作1080。在操作1080中,神经网络装置可将第一层的输出特征图存储在片上缓冲存储器中。当在操作S1073中确定第一层的输出特征图的大小超过片上缓冲存储器的大小时,神经网络装置可执行操作S1072。在操作S1072中,神经网络装置可将第一层的输出特征图存储在辅助缓冲存储器中。

[0146] 在操作1090中,神经网络装置可重新使用第一层的输出特征图作为第二层($n=2$)的输入特征图。当第一层的输出特征图被重新用作第二层的输入特征图时,后续操作中的下一层可对应于“ $n=n+1$ ”。在操作1100中,神经网络装置可将第二层的输入特征图发送到计算电路以执行神经网络运算并输出第二层的输出特征图。

[0147] 在操作1110中,神经网络装置确定第二层是否是最后层。当确定第二层是最后层时,神经网络装置可执行操作1120。在操作1120中,神经网络装置可将输出特征图存储在外部存储器中。当确定第二层不是最后层时,神经网络装置可执行操作1070以基于第二层的输入特征图和输出特征图的总大小来确定是否在兼容模式下操作。

[0148] 随后,在重复操作1090时,对下一层执行相同的前述处理,并且当到达最后层时,神经网络装置可将神经网络的最后层的输出特征图存储在外部存储器中并终止神经网络装置的操作。

[0149] 执行在本申请中描述的操作的图1至图10中的神经网络装置、神经网络装置300、神经网络装置500、神经网络装置830、片上缓冲存储器310、片上缓冲存储器510、计算电路320、计算电路520、控制器330、控制器530、外部存储器390、外部存储器590、辅助缓冲存储器550、CPU 810、RAM 820、存储器840、传感器模块850、发送/接收模块860由硬件组件实现,硬件组件被配置为执行由硬件组件执行的在本申请中描述的操作。可用于执行在本申请中描述的操作的硬件组件的示例在适当的情况下包括:控制器、传感器、生成器、驱动器、存储器、比较器、算术逻辑单元、加法器、减法器、乘法器、除法器、积分器和被配置为执行在本申请中描述的操作的任何其它电子组件。在其它示例中,执行在本申请中描述的操作的硬件

组件中的一个或多个通过计算硬件(例如,通过一个或多个处理器或计算机)来实现。处理器或计算机可通过一个或多个处理元件(诸如,逻辑门阵列、控制器和算术逻辑单元、数字信号处理器、微型计算机、可编程逻辑控制器、现场可编程门阵列、可编程逻辑阵列、微处理器或被配置为以限定的方式响应并执行指令以实现期望的结果的任何其它装置或装置的组合)来实现。在一个示例中,处理器或计算机包括或连接到存储由处理器或计算机执行的指令或软件的一个或多个存储器。由处理器或计算机实现的硬件组件可执行用于执行在本申请中描述的操作的指令或软件(诸如,操作系统(OS)和在OS上运行的一个或多个软件应用)。硬件组件也可响应于指令或软件的执行来访问、操控、处理、创建和存储数据。为了简明起见,单数术语“处理器”或“计算机”可用在本申请中描述的示例的描述中,但是在其它示例中,多个处理器或计算机可被使用,或者处理器或计算机可包括多个处理元件或多种类型的处理元件或二者。例如,单个硬件组件或者两个或更多个硬件组件可通过单个处理器、或者两个或更多个处理器、或者处理器和控制器来实现。一个或多个硬件组件可通过一个或多个处理器、或者处理器和控制器来实现,并且一个或多个其它硬件组件可通过一个或多个其它处理器、或者另外的处理器和另外的控制器来实现。一个或多个处理器、或者处理器和控制器可实现单个硬件组件或者两个或更多个硬件组件。硬件组件可具有不同的处理配置中的任何一个或多个,不同的处理配置的示例包括:单个处理器、独立处理器、并行处理器、单指令单数据(SISD)多处理、单指令多数据(SIMD)多处理、多指令单数据(MISD)多处理以及多指令多数据(MIMD)多处理。

[0150] 图1至图10中示出的执行在本申请中描述的操作的方法通过计算硬件(例如,通过一个或多个处理器或计算机)来执行,计算硬件被如上所述地实现为执行指令或软件,以执行在本申请中描述的由所述方法执行的操作。例如,单个操作或者两个或更多个操作可通过单个处理器、或者两个或更多个处理器、或者处理器和控制器来执行。一个或多个操作可通过一个或多个处理器、或者处理器和控制器来执行,并且一个或多个其它操作可通过一个或多个其它处理器、或者另外的处理器和另外的控制器来执行。一个或多个处理器、或者处理器和控制器可执行单个操作或者两个或更多个操作。

[0151] 用于控制计算硬件(例如,一个或多个处理器或计算机)以实现硬件组件并执行如上所述的方法的指令或软件可被编写为计算机程序、代码段、指令或它们的任意组合,以单独地或共同地指示或配置一个或多个处理器或计算机如机器或专用计算机那样进行操作,以执行由如上所述的硬件组件和方法执行的操作。在一个示例中,指令或软件包括由一个或多个处理器或计算机直接执行的机器代码(诸如,由编译器产生的机器代码)。在另一示例中,指令或软件包括由一个或多个处理器或计算机使用解释器执行的高级代码。可基于附图中示出的框图和流程图以及说明书中的相应描述,使用任何编程语言编写指令或软件,附图中示出的框图和流程图以及说明书中的相应描述公开了用于执行由如上所述的硬件组件和方法执行的操作的算法。

[0152] 用于控制计算硬件(例如,一个或多个处理器或计算机)以实现硬件组件并执行如上所述的方法的指令或软件以及任何相关联的数据、数据文件和数据结构可被记录、存储或固定在一个或多个非暂时性计算机可读存储介质中或一个或多个非暂时性计算机可读存储介质上。非暂时性计算机可读存储介质的示例包括:只读存储器(ROM)、随机存取存储器(RAM)、闪存、CD-ROM、CD-R、CD+R、CD-RW、CD+RW、DVD-ROM、DVD-R、DVD+R、DVD-RW、DVD+RW、

DVD-RAM、BD-ROM、BD-R、BD-R LTH、BD-RE、磁带、软盘、磁光数据存储装置、光学数据存储装置、硬盘、固态盘、以及任何其它装置,任何其它装置被配置为以非暂时性方式存储指令或软件以及任何相关联的数据、数据文件和数据结构并将指令或软件以及任何相关联的数据、数据文件和数据结构提供给一个或多个处理器或计算机,使得一个或多个处理器和计算机可执行指令。在一个示例中,指令或软件以及任何相关联的数据、数据文件和数据结构分布在联网的计算机系统上,使得指令和软件以及任何相关联的数据、数据文件和数据结构通过一个或多个处理器或计算机以分布式的方式被存储、访问和执行。

[0153] 虽然本公开包括特定的示例,但是在理解本申请的公开之后将清楚,在不脱离权利要求及它们的等同物的精神和范围的情况下,可在这些示例中进行形式和细节上的各种改变。在此描述的示例将被认为仅是描述性的,而不是出于限制的目的。每个示例中的特征或方面的描述将被认为可适用于其它示例中的类似特征或方面。如果描述的技术以不同的次序被执行,和/或如果描述的系统、架构、装置或电路中的组件以不同的方式被组合,和/或由其它组件或它们的等同物替换或补充,则可实现合适的结果。因此,公开的范围不是由具体实施方式限定,而是由权利要求及它们的等同物限定,并且在权利要求及它们的等同物的范围内的所有变化应被解释为包括在公开中。

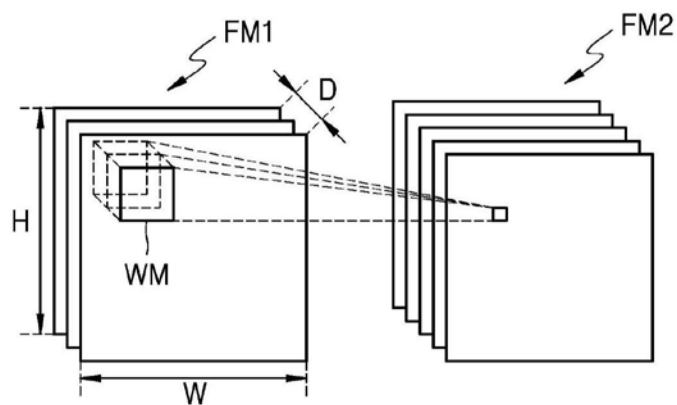


图1

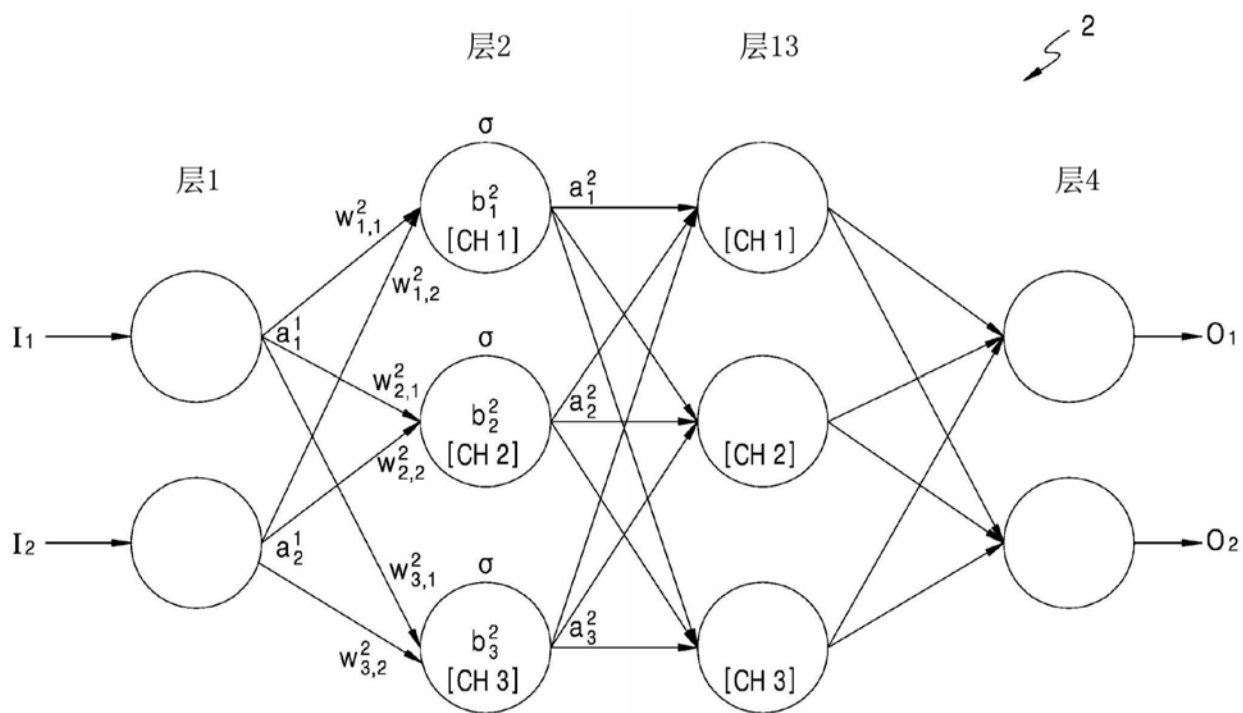


图2

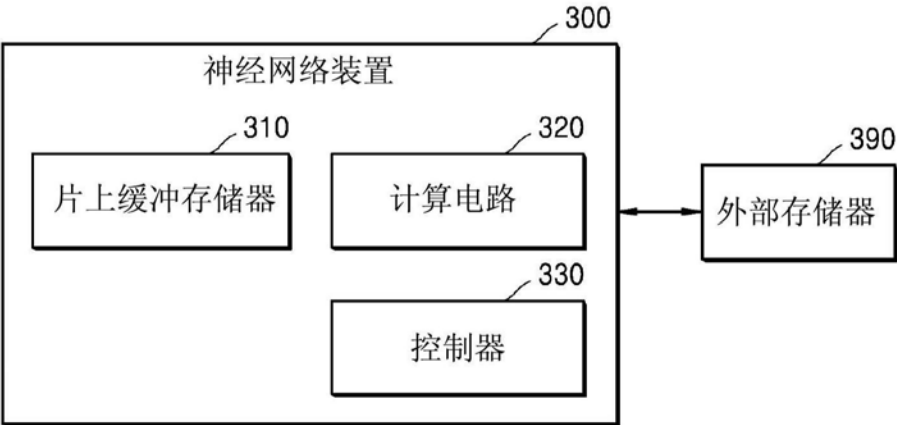


图3

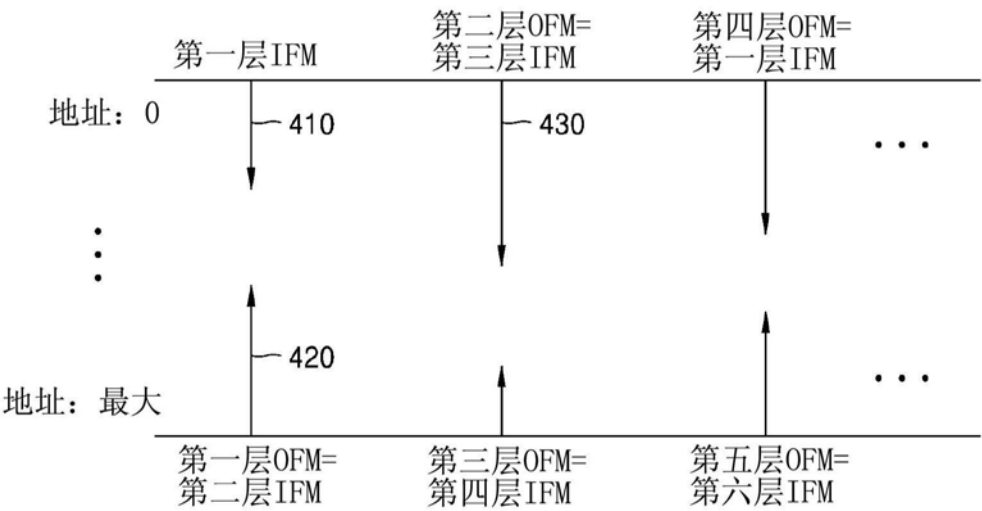


图4

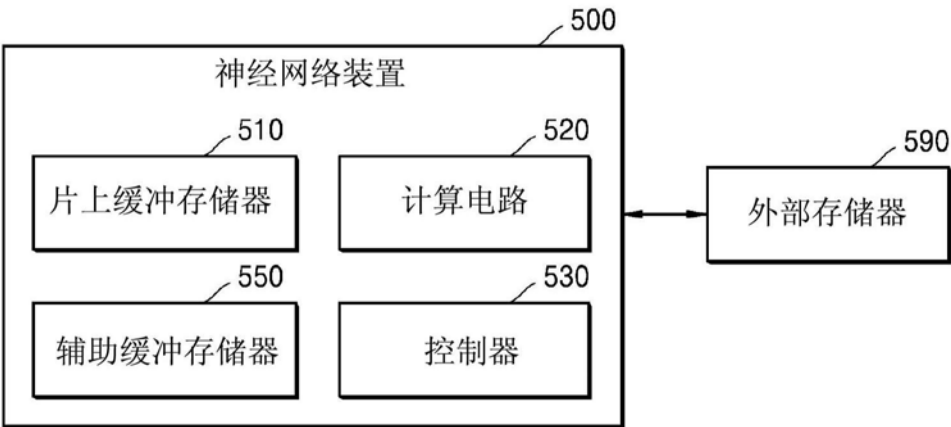


图5

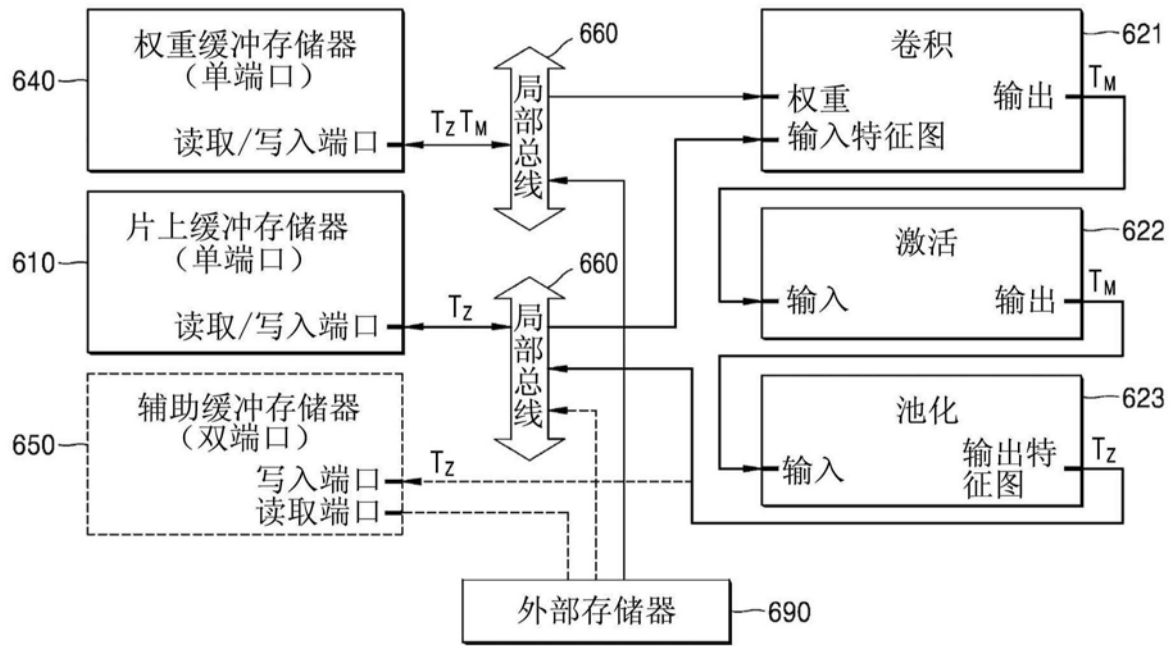


图6

参数	定义
M	输出特征图 (OFM) 中的特征的数量
R,C	OFM的行的数量 (R) 和列的数量 (C)
Z	输入特征图 (IFM) 中的特征的数量
K	二维卷积的滤波器在每个维度中的大小
S	二维卷积在每个维度中的步长
P	二维池化在每个维度中的步长
Q	二维池化核在每个维度中的大小

```

for (m = 0; m < M; m += TM){                                     // M-loop
  WeightBias DRAM read();
  for (rr = 0; rr < R; rr += P){                                   // RR-loop
    for (cc = 0; cc < C; cc += P){                                 // CC-loop
700 — if (IFM_offchip_mode) IFM_DRAM_read();
      for (r = rr; r < min(rr + P, R); r++){                     // R-loop
        for (c = cc; c < min(cc + P, C); c++){                   // C-loop
          Initialize OFM with bias();

          for (i = 0; i < K; i++){                                // K-loop #1
            for (j = 0; j < K; j++){                               // K-loop #2
              for (z = 0; z < Z; z += TZ){                       // Z-loop
                OFM[m : m + TM][r][c] +=
                  W[m : m + TM][z : z + TZ][i][j] ·
                  IFM[z : z + TZ][S · r + i][S · c + j];
              }}}
          Activation();
          Pooling();
        }
      }
710 — if (OFM_offchip_mode) OFM_DRAM_write();
    }
  }
}

```

// R-loop 结束
// M-loop 结束

图7

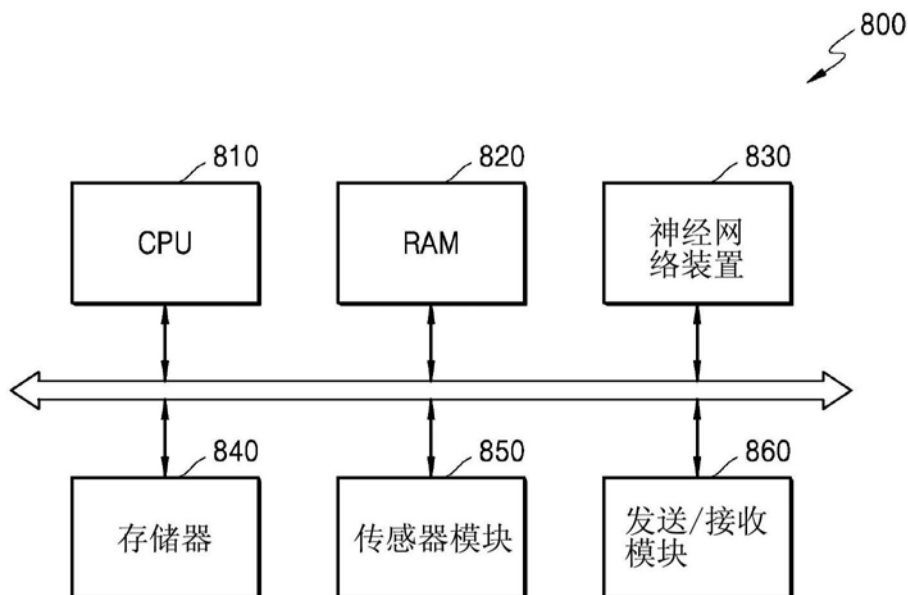


图8

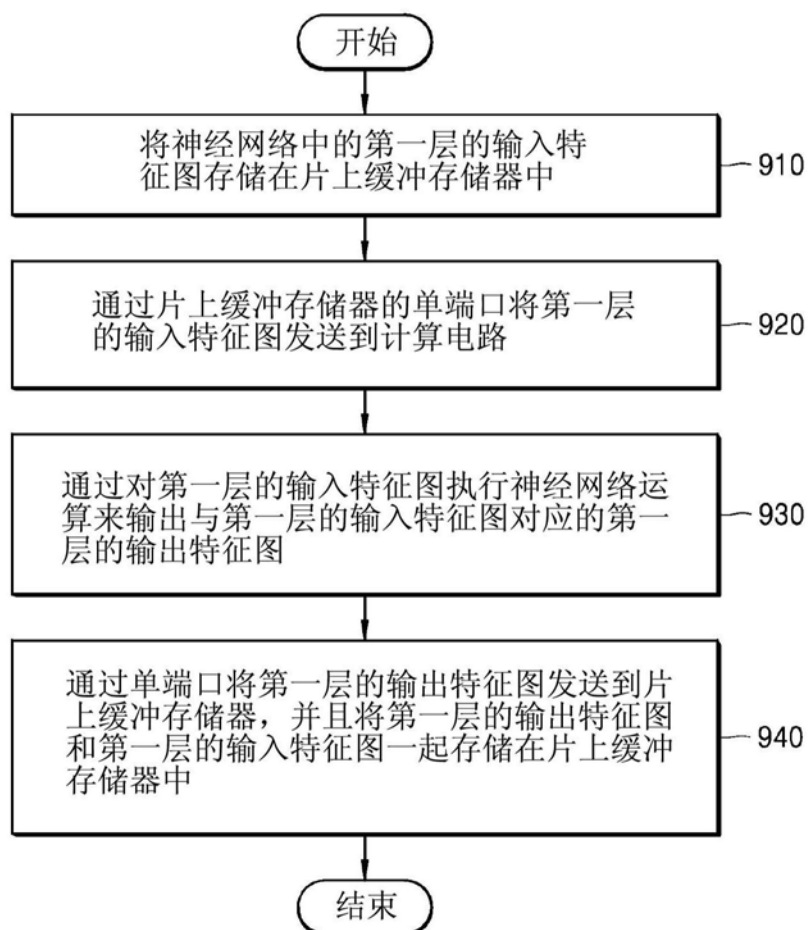


图9

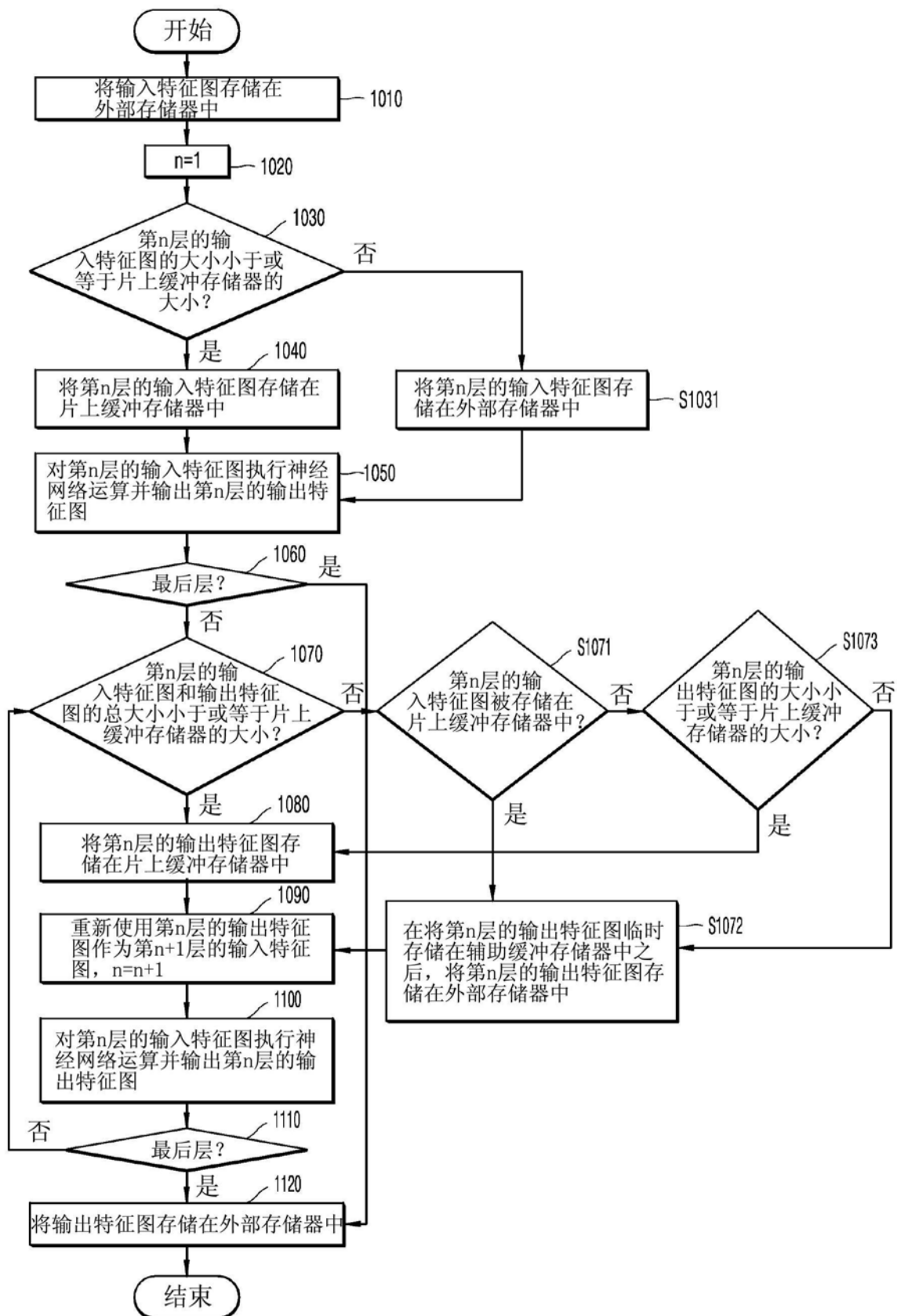


图10