(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 108256644 B
(45)授权公告日 2021.06.22

(21)申请号 201810012336.8

(22)申请日 2018.01.05

(65)同一申请的已公布的文献号
申请公布号 CN 108256644 A

(43)申请公布日 2018.07.06

(73)专利权人 上海兆芯集成电路有限公司
地址 201203 上海市浦东新区张江高科技园区金科路2537号301室

(72)发明人 李晓阳 陈静

(74)专利代理机构 北京汇泽知识产权代理有限公司 11228
代理人 张瑾

(51)Int.Cl.
G06N 3/063(2006.01)
G06F 15/80(2006.01)

(56)对比文件
CN 103019656 A,2013.04.03
CN 106875012 A,2017.06.20
CN 107256424 A,2017.10.17
CN 103019656 A,2013.04.03
王敏等.循环神经网络语言模型定点化优化算法研究.《软件导刊》.2017,第16卷(第2期),
M.Kemal Ciliz等.Recurrent Neural Network Model on an SIMD Parallel Architecture.《Proceedings of 1993 International Joint Conference on Neural Networks》.2002,

审查员 刘娜

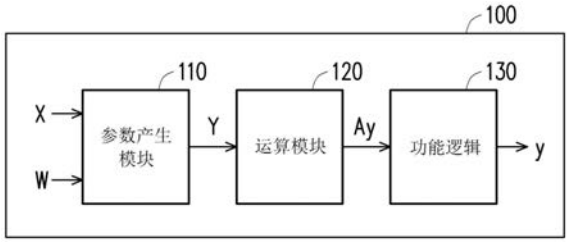
权利要求书3页 说明书11页 附图7页

(54)发明名称

微处理器电路以及执行神经网络运算的方法

(57)摘要

本发明提出一种微处理器电路以及一种执行神经网络运算的方法。所述微处理器电路适用于执行神经网络运算。所述微处理器电路包括参数产生模块、运算模块以及比较逻辑。所述参数产生模块并行接收所述神经网络运算的多个输入参数以及多个权重参数。所述参数产生模块依据所述多个输入参数以及所述多个权重参数来并行产生多个子输出参数。所述运算模块并行接收所述多个子输出参数。所述运算模块加总所述多个子输出参数,以产生加总参数。所述比较逻辑接收所述加总参数。所述比较逻辑依据所述加总参数执行比较运算,以产生所述神经网络运算的输出参数。



CN 108256644 B

1.一种支持单指令流多数据流架构的微处理器电路,适用于执行神经网络运算,其特征在于,包括:

参数产生模块,用以并行接收所述神经网络运算的多个输入参数以及多个权重参数,并且所述参数产生模块依据所述多个输入参数以及所述多个权重参数来并行产生多个子输出参数,其中,所述参数产生模块依据所述多个输入参数及所述多个权重参数的取值范围对所述多个输入参数以及所述多个权重参数采用相同方式进行编码,以产生多个编码后的输入参数以及多个编码后的权重参数,并且所述参数产生模块依据所述多个编码后的输入参数以及所述多个编码后的权重参数来产生所述多个子输出参数;

运算模块,耦接所述参数产生模块,并且用以并行接收所述多个子输出参数,其中所述运算模块加总所述多个子输出参数,以产生加总参数;以及

比较逻辑,耦接所述运算模块,并且用以接收所述加总参数,其中所述比较逻辑依据所述加总参数执行比较运算,以产生所述神经网络运算的输出参数;

其中,所述神经网络运算为二值神经网络运算、三值神经网络运算、二值权重网络运算或三值权重网络运算。

2.根据权利要求1所述的微处理器电路,其特征在于,所述比较逻辑依据所述加总参数来比较所述多个子输出参数的第一数值类型的数量以及第二数值类型的数量,以决定所述输出参数。

3.根据权利要求1所述的微处理器电路,其特征在于,若所述多个输入参数以及所述多个权重参数的取值范围分别包括两种数值类型,则所述参数产生模块采用第一编码方式对所述多个输入参数以及所述多个权重参数进行编码,并且所述参数产生模块依据所述多个编码后的输入参数以及所述多个编码后的权重参数来透过第一查找表或第一逻辑电路来产生所述多个子输出参数。

4.根据权利要求1所述的微处理器电路,其特征在于,若所述多个输入参数以及所述多个权重参数的取值范围分别包括三种数值类型,则所述参数产生模块采用第二编码方式对所述多个输入参数以及所述多个权重参数编码,并且所述参数产生模块依据所述多个编码后的输入参数以及所述多个编码后的权重参数来透过第二查找表或第二逻辑电路来产生所述多个子输出参数。

5.根据权利要求4所述的微处理器电路,其特征在于,所述运算模块包括:

第一子运算模块,用以加总所述多个子输出参数的第一位的数值,以产生第一加总参数;以及

第二子运算模块,用以加总所述多个子输出参数的第二位的数值,以产生第二加总参数,

其中所述比较逻辑比较所述第一加总参数与所述第二加总参数,以决定所述输出参数。

6.根据权利要求4所述的微处理器电路,其特征在于,所述第二逻辑电路包括:

第一子逻辑电路,用以依据编码后的所述神经网络运算的所述多个输入参数以及编码后的所述多个权重参数来产生所述多个子输出参数的第一位;以及

第二子逻辑电路,用以依据编码后的所述神经网络运算的所述多个输入参数以及编码后的所述多个权重参数来产生所述多个子输出参数的第二位。

7.根据权利要求4所述的微处理器电路，其特征在于，所述运算模块计算所述多个子输出参数的第一位为第一数值的个数以及所述多个子输出参数的第二位为所述第一数值的个数。

8.根据权利要求4所述的微处理器电路，其特征在于，所述比较逻辑比较所述多个子输出参数的第一位为第一数值的个数以及所述多个子输出参数的第二位为所述第一数值的个数来决定所述输出参数。

9.根据权利要求1所述的微处理器电路，其特征在于，所述微处理器电路执行微指令以完成所述神经网络运算，所述微指令的源操作数包括上述多个输入参数及上述多个权重参数，并且所述微指令的目的操作数包括所述神经网络运算的所述输出参数。

10.根据权利要求1所述的微处理器电路，其特征在于，每一所述多个输入参数的位宽等于每一所述多个权重参数的位宽，并且所述微处理器电路的位宽大于所述多个输入参数及所述多个权重参数的位宽总和。

11.一种执行神经网络运算的方法，适用于支持单指令流多数据流架构的微处理器电路，其特征在于，所述微处理器电路包括参数产生模块、运算模块以及比较逻辑，所述方法包括：

透过参数产生模块并行接收所述神经网络运算的多个输入参数以及多个权重参数，并且依据所述多个输入参数以及所述多个权重参数来并行产生多个子输出参数；

透过运算模块并行接收所述多个子输出参数，并且加总所述多个子输出参数，以产生加总参数；以及

透过比较逻辑接收所述加总参数，并且依据所述加总参数执行比较运算，以产生所述神经网络运算的输出参数；

其中，依据所述多个输入参数以及所述多个权重参数来并行产生多个子输出参数的步骤包括：透过所述参数产生模块依据所述多个输入参数及所述多个权重参数的取值范围对所述多个输入参数以及所述多个权重参数采用相同方式进行编码，以产生多个编码后的输入参数以及多个编码后的权重参数；以及透过所述参数产生模块依据所述多个编码后的输入参数以及所述多个编码后的权重参数来产生所述多个子输出参数；

其中，所述神经网络运算为二值神经网络运算、三值神经网络运算、二值权重网络运算或三值权重网络运算。

12.根据权利要求11所述的神经网络运算的方法，其特征在于，依据所述加总参数执行比较运算，以产生所述神经网络运算的输出参数的步骤包括：

透过所述比较逻辑依据所述加总参数来比较所述多个子输出参数的第一数值类型的数量以及第二数值类型的数量，以决定所述输出参数。

13.根据权利要求11所述的神经网络运算的方法，其特征在于，若所述多个输入参数以及所述多个权重参数的取值范围分别包括两种数值类型，则所述参数产生模块采用第一编码方式对所述多个输入参数以及所述多个权重参数进行编码，并且依据所述多个输入参数以及所述多个权重参数来并行产生所述多个子输出参数的步骤包括：

透过所述参数产生模块依据所述多个编码后的输入参数以及所述多个编码后的权重参数来透过第一查找表或第一逻辑电路来产生所述多个子输出参数。

14.根据权利要求11所述的神经网络运算的方法，其特征在于，若所述多个输入参数以

及所述多个权重参数的取值范围分别包括三种数值类型，则所述参数产生模块采用第二编码方式对所述多个输入参数以及所述多个权重参数编码，并且依据所述多个输入参数以及所述多个权重参数来并行产生所述多个子输出参数的步骤包括：

透过所述参数产生模块依据所述多个编码后的输入参数以及所述多个编码后的权重参数来透过第二查找表或第二逻辑电路来产生所述多个子输出参数。

15.根据权利要求14所述的神经网络运算的方法，其特征在于，所述运算模块包括第一子运算模块以及第二子运算模块，并且加总所述多个子输出参数，以产生加总参数的步骤包括：

透过所述第一子运算模块加总所述多个子输出参数的第一位的数值，以产生第一加总参数；以及

透过所述第二子运算模块加总所述多个子输出参数的第二位的数值，以产生第二加总参数，

其中依据所述加总参数执行比较运算，以产生所述神经网络运算的输出参数的步骤包括：

透过所述比较逻辑比较所述第一加总参数与所述第二加总参数，以决定所述输出参数。

16.根据权利要求14所述的神经网络运算的方法，其特征在于，所述第二逻辑电路包括第一子逻辑电路以及第二子逻辑电路，并且透过第二查找表或第二逻辑电路来产生所述多个子输出参数的步骤包括：

透过所述第一子逻辑电路依据编码后的所述神经网络运算的所述多个输入参数以及编码后的所述多个权重参数来产生所述多个子输出参数的第一位；以及

透过所述第二子逻辑电路依据编码后的所述神经网络运算的所述多个输入参数以及编码后的所述多个权重参数来产生所述多个子输出参数的第二位。

17.根据权利要求14所述的神经网络运算的方法，其特征在于，所述运算模块计算所述多个子输出参数的第一位为第一数值的个数以及所述多个子输出参数的第二位为所述第一数值的个数。

18.根据权利要求14所述的神经网络运算的方法，所述比较逻辑比较所述多个子输出参数的第一位为第一数值的个数以及所述多个子输出参数的第二位为所述第一数值的个数来决定所述输出参数。

微处理器电路以及执行神经网络运算的方法

技术领域

[0001] 本发明是有关于一种单指令流多数据流(Single Instruction Multiple Data，SIMD)架构的应用，且特别是有关于一种应用此架构的微处理器电路以及执行神经网络运算的方法。

背景技术

[0002] 一般而言，传统的处理器执行神经网络运算(Neural network operation)需要利用大量的储存空间。在一般的情况下，神经网络运算将在静态随机存取存储器(Static Random Access Memory，SRAM)中占用大量的储存空间，或是在静态随机存取存储器与动态随机存取存储器(Dynamic Random Access Memory，DRAM)之间进行大量的数据交换。并且，当处理器例如执行二值神经网络

[0003] (Binary Neural Network，BNN)运算或三值神经网络(Ternary Neural Network，TNN)运算时，输出参数$y=\sum_{i=1}^{n}w_i x_i$，其中w_i为一位(1-bit)或二位(2-bit)的权重参数，x_i为与w_i位宽相等的输入数据，y为输出参数，其中一位(1-bit)或二位(2-bit)的权重参数w_i以及输入数据x_i还需配合八位(8-bit)单指令流多数据流(SIMD)通道(Lane)进行处理。当处理器例如执行二值权重网络(Binary Weight Network，BWN)运算或三值权重网络(Ternary Weight Network，TWN)运算时，输出参数$y=\sum_{i=1}^{n}w_i x_i$，其中w_i为一位(1-bit)或二位(2-bit)权重参数，x_i为八位(8-bit)输入数据，y为输出参数，其中一位(1-bit)或二位(2-bit)的权重参数w_i需配合八位(8-bit)单指令流多数据流(SIMD)通道(Lane)进行处理。因此，传统的处理器执行神经网络运算的方式会造成运算资源的浪费。有鉴于此，以下将提出几个解决方案。

发明内容

[0004] 本发明提供一种微处理器电路以及执行神经网络运算(Neural network operation)的方法，可有效节省神经网络运算的运算资源。

[0005] 本发明的一种微处理器电路适用于执行神经网络运算。所述微处理器电路包括参数产生模块、运算模块以及比较逻辑。所述参数产生模块用以并行接收所述神经网络运算的多个输入参数以及多个权重参数。所述参数产生模块依据所述多个输入参数以及所述多个权重参数来并行产生多个子输出参数。所述运算模块耦接所述参数产生模块。所述运算模块用以并行接收所述多个子输出参数。所述运算模块加总所述多个子输出参数，以产生加总参数。所述比较逻辑耦接所述运算模块。所述比较逻辑用以接收所述加总参数。所述比较逻辑依据所述加总参数执行比较运算，以产生所述神经网络运算的输出参数。

[0006] 本发明的一种执行神经网络运算的方法适用于微处理器电路。所述微处理器电路包括参数产生模块、运算模块以及比较逻辑。所述方法包括以下步骤：透过所述参数产生模

块并行接收所述神经网络运算的多个输入参数以及多个权重参数，并且依据所述多个输入参数以及所述多个权重参数来产生多个子输出参数。透过所述运算模块并行接收所述多个子输出参数，并且加总所述多个子输出参数，以产生加总参数。透过所述比较逻辑并行接收所述加总参数，并且依据所述加总参数执行比较运算，以产生所述神经网络运算的输出参数。

[0007] 基于上述，本发明的微处理器电路以及执行神经网络运算的方法可应用单指令流多数据流(Single Instruction Multiple Data，SIMD)架构，并且当执行二值神经网络(Binary Neural Network，BNN)运算或三值神经网络(Ternary Neural Network，TNN)运算时，可实现非常高的并行性(Parallelism)，以有效节省神经网络运算的运算资源。

[0008] 为让本发明的上述特征和优点能更明显易懂，下文特举实施例，并配合所附图式作详细说明如下。

附图说明

[0009] 图1是依照本发明的一实施例的微处理器电路的示意图。

[0010] 图2是依照本发明的第一实施例的微处理器电路的示意图。

[0011] 图3是依照本发明的第一实施例的执行神经网络运算的方法的流程图。

[0012] 图4是依照本发明的第二实施例的微处理器电路的示意图。

[0013] 图5是依照本发明的第二实施例的执行神经网络运算的方法的流程图。

[0014] 图6是依照本发明的第三实施例的微处理器电路的示意图。

[0015] 图7是依照本发明的第三实施例的执行神经网络运算的方法的流程图。

具体实施方式

[0016] 为了使本发明的内容可以被更容易明了，以下特举实施例做为本发明确实能够据以实施的范例。另外，凡可能之处，在图式及实施方式中使用相同标号的组件/构件/步骤，代表相同或类似部件。

[0017] 图1是依照本发明的一实施例的微处理器电路的示意图。参考图1，微处理器电路100包括参数产生模块110、运算模块120以及功能逻辑(Logic)130。参数产生模块110耦接运算模块120。运算模块120耦接功能逻辑130。在本实施例中，微处理器电路100例如是一种通用处理器(General Purpose Processor，GPP)，并且微处理器电路100整合一个或多个处理单元来执行对应的运算工作。在本实施例中，微处理器电路100是以单指令流多数据流(Single Instruction Multiple Data，SIMD)架构来执行神经网络运算(Neural Network Computing)的微指令(micro-instruction或μop)以完成对应的神经网络运算，微处理器电路100为包含于处理器核心(core)的执行单元(execution unit)中的硬件电路。值得注意的是，本实施例所提及的微指令是指微处理器电路100所属的微处理器能直接执行的指令。所述微处理器可例如是复杂指令集计算机(CISC)指令集架构的超标量乱序执行处理器、精简指令集计算机(RISC)指令集架构的处理器、为执行神经网络运算而设计的专有架构处理器或是其它架构的处理器。

[0018] 在本实施例中，当微处理器电路100执行神经网络运算工作时，参数产生模块110并行接收神经网络运算的输入数据X以及权重数据W，并且参数产生模块110依据输入数据X

以及权重数据W来产生子输出数据Y至运算模块120。在本实施例中，输入数据X、权重数据W以及子输出数据Y可分别包括多个数值。在本实施例中，输入数据X包括多个并行的输入参数(x_1、x_2～x_n)。权重数据W包括多个并行的权重参数(w_1、w_2～w_n)。子输出数据Y包括多个并行的子输出参数(y_1、y_2～y_n)。在本实施例中，参数产生模块110例如透过查找表(Look-up table)或特定逻辑电路来依据并行的多个输入参数(x_1、x_2～x_n)以及并行的多个权重参数(w_1、w_2～w_n)，来并行产生多个并行的子输出参数(y_1、y_2～y_n)。

[0019] 在本实施例中，运算模块120并行接收参数产生模块110提供的包括多个子输出参数(y_1、y_2～y_n)的子输出数据Y，并且对多个子输出参数(y_1、y_2～y_n)进行运算。运算模块120对具有多个数值的子输出数据Y进行并行运算(parallel computing)。举例来说，运算模块120包括一个或多个加法器(Adder)，以透过加法器来加总子输出参数Y的多个数值(y_1、y_2～y_n)，并且产生加总参数Ay。也就是说，在本实施例中，运算模块120对具有多个数值的子输出参数(y_1、y_2～y_n)并行执行加法运算，以有效率地产生加总参数Ay。

[0020] 在本实施例中，功能逻辑130为一个或多个运算逻辑组成的硬件架构。功能逻辑130例如是比较逻辑(compare logic)或截位逻辑(truncation logic)等诸如此类的逻辑。功能逻辑130接收运算模块120提供的加总参数Ay，其中加总参数Ay可以为单一笔数值。功能逻辑130对加总参数Ay进行特定功能运算，以产生神经网络运算的输出参数y。也就是说，本实施例的微处理器电路100透过直接执行一条微指令(micro-instruction或μop)，即可有效率地产生神经网络运算的输出参数y。

[0021] 举例而言，上述的微指令可例如是“MAC,Dst,Scr1,Scr2,size 1,size 2,size3”。在一实施例中，微处理器电路100可执行单一的一条这样的微指令即可完成特定神经网络运算$y=\sum_{i=1}^{n} w_i x_i$，其中字段MAC为操作码。微处理器电路100识别此操作码以得知由自己来处理此微指令。需注意的是，微处理器电路100仅为微处理器的执行单元(execution unit)的一部分，所述微处理器的执行单元可进一步包含执行其它类型指令的其它执行电路。在上述的微指令中，字段“Scr1”用于指示微指令的一源操作数(source operand)，并且包括上述多个输入参数(x_1、x_2～x_n)。字段“Scr2”用于指示微指令的另一源操作数，并且包括上述多个权重参数(w_1、w_2～w_n)。字段“Dst”用于指示微指令的目的操作数(destination operand)，并且取得上述神经网络运算的输出参数y。字段“size 1”用于指示每个输入参数(x_1、x_2～x_n)的位宽(bit width)。字段“size 2”用于指示每个权重参数(w_1、w_2～w_n)的位宽。字段“size 3”用于指示输出参数y的位宽。然而，上述的微指令的格式仅为本发明的其中一种实施范例，本发明并不限于此。

[0022] 进一步来说，本实施例的微处理器电路100的实施架构可例如适用于执行二值神经网络(Binary Neural Network,BNN)运算、三值神经网络(Ternary Neural Network,TNN)运算、二值权重网络(Binary Weight Network,BWN)运算以及三值权重网络(Ternary Weight Network,TWN)运算等诸如此类的神经网络运算。在BNN运算和TNN运算中，每个输入参数(x_1、x_2～x_n)的位宽等于每个权重参数(w_1、w_2～w_n)的位宽，例如为2bit，而输出参数y的位宽例如为8bit。在BWN运算和TWN运算中，每个权重参数w_i的位宽例如为2bit，并且每个输入参数(x_1、x_2～x_n)的位宽大于每个权重参数(w_1、w_2～w_n)的位宽，例如皆为8bit。也就是说，BNN运算和TNN运算的两个源操作数的位宽“size 1”和“size 2”相等。BWN运算和TWN运算的

两个源操作数的位宽“size 1”和“size 2”不相等，并且“size 1”大于“size 2”。对此，本发明的微处理器电路100的位宽(例如为256/128bit)大于所述多个输入参数(x_1、x_2～x_n)及所述多个权重参数(w_1、w_2～w_n)的位宽总和。

[0023] 在本实施例中，微处理器电路100并行接收神经网络运算的多个较短的输入参数(x_1、x_2～x_n)及多个权重参数(w_1、w_2～w_n)，以执行运算$y=\sum_{i=1}^{n} w_i x_i$，

[0024] 以使微处理器电路100可适应神经网络运算具有数据量大且数据位宽短的特点，并且实现并行运算。因此，本实施例的微处理器电路100可有效节省运算周期，并且提高运算效率。为了更进一步说明本发明的微处理器电路的多种的实施架构，以下各范例实施例的微处理器电路将分别搭配执行对应的神经网络运算方法来说明之。

[0025] 图2是依照本发明的第一实施例的微处理器电路的示意图。图3是依照本发明的第一实施例的执行神经网络运算的方法的流程图。参考图2以及图3，微处理器电路200包括参数产生模块210、运算模块220以及比较逻辑230。在本实施例中，微处理器电路200适用于执行BNN运算，并且微处理器电路200欲实现的BNN运算如以下公式(1)。在本实施例中，输入参数x_1、x_2～x_n、权重参数w_1、w_2～w_n及输出参数y的取值范围为{-1,1}。并且，依据BNN运算的算法规定，若输入参数x_1、x_2～x_n以及权重参数w_1、w_2～w_n的求和结果大于或等于0(y≥0)，则y＝1。反之，则y＝-1。

[0026] $$y=\sum_{i=1}^{n} w_i x_i = w_1x_1 + w_2x_2 + w_3x_3 + \cdots + w_nx_n \ , \ i>1 \quad \text{.....公式 (1)}$$

[0027] 首先，在步骤S310中，微处理器电路200透过参数产生模块210并行接收神经网络运算的多个输入参数x_1、x_2～x_n以及多个权重参数w_1、w_2～w_n。参数产生模块210依据输入参数x_1、x_2～x_n以及权重参数w_1、w_2～w_n来产生多个子输出参数y_1、y_2～y_n，其中n为大于0的正整数。在本实施例中，输入参数x_1、x_2～x_n以及权重参数w_1、w_2～w_n的取值范围为{-1,1}。在本实施例中，参数产生模块210预先对输入参数x_1、x_2～x_n以及权重参数w_1、w_2～w_n进行编码，其中编码结果例如以下表1。并且，在本实施例中，编码后的输入参数x_1、x_2～x_n、编码后的权重参数w_1、w_2～w_n以及对应的子输出参数y_1、y_2～y_n的数值如以下表2(第一查找表)，其中i∈{1,2,3,…,n}。

[0028]

w_i,x_i	w_i,x_i(编码后)
-1	0
1	1

[0029] 表1

[0030]

w_i（编码后）	x_i（编码后）	y_i
0	0	1
0	1	0
1	0	0
1	1	1

[0031] 表2

[0032] 在本实施例中，参数产生模块210依据上述表2来并行取得所有的子输出参数y_1、y_2～y_n，但本发明并不限于此。在一实施例中，参数产生模块210可以采用其它编码方式来产生子输出参数y_1、y_2～y_n，例如1编码为1，-1编码为0。在一实施例中，参数产生模块210也可依据以下公式(2)对应的数字逻辑电路(第一逻辑电路)来取得对应的子输出参数y_1、y_2～y_n。如图2所示，参数产生模块210可包括第一逻辑电路。在本实施例中，以下公式(2)当中的符号“～”以及符号“^”为操作数(operator)的符号，其中可分别代表“取反(INV)”以及“异或(XOR)”。注意本发明不限于此，例如在将数值为1的输入参数x_i或权重参数w_i编码为1，将数值为-1的输入参数x_i或权重参数w_i编码为0的实施方式中，可采用另一公式$y_i=w_i \hat{} x_i$对应的数字逻辑电路来取得对应的子输出参数y_1、y_2～y_n。

[0033] $y_i=\sim(w_i \hat{} x_i)$ .....公式(2)

[0034] 接着，在步骤S320中，微处理器电路200透过运算模块220并行接收子输出参数y_1、y_2～y_n，并且加总子输出参数y_1、y_2～y_n，以产生加总参数Ay。在本实施例中，运算模块220包括多个加法器221。运算模块220对子输出参数y_1、y_2～y_n同时进行加法运算。举例来说，加法器221的第一层对子输出参数y_1、y_2～y_n的每四个数据进行相加的运算，并且提供至这些加法器221的第二层。以此类推，这些加法器221的最后一层可输出加总参数Ay。在本实施例中，每个加法器221可由基本的4:2进位保留加法器(Carry Save Adder，CSA)实现，每个CSA加法器包括4个输入和2个输出(包括总和Sum和进位Carry)。然而，运算模块220的这些加法器221的数量及层数可依据子输出参数y_1、y_2～y_n的数量来对应设计，本发明的运算模块220的这些加法器221的配置方式不限于图2所示。

[0035] 在本实施例中，由于BNN运算的算法中输入参数x_1、x_2～x_n以及权重参数w_1、w_2～w_n的取值范围为{-1,1}，即取值仅两种数值“-1”和“1”。因此，参数产生模块210采用1-bit编码方式对输入参数x_1、x_2～x_n以及权重参数w_1、w_2～w_n进行编码后产生的编码后数值为“0”或“1”。换句话说，运算模块220经由多个加法器221运算输出参数y_1、y_2～y_n，以取得加总参数Ay，其目的在于取得子输出参数y_1、y_2～y_n当中的第一数值类型(“1”)的数量。加总参数Ay即对应上述公式(1)的乘加求和运算的结果。

[0036] 最后，在步骤S330中，微处理器电路200透过比较逻辑230接收加总参数Ay，并且比较逻辑230依据加总参数Ay执行比较运算，以产生神经网络运算的输出参数y。在本实施例

中，比较逻辑230依据加总参数Ay来判断子输出参数y_1、y_2～y_n当中的第一数值类型（“1”）的数量以及第二数值类型（“0”）的数量，以决定输出参数y为“1”或“0”。举例来说，比较逻辑230可执行如以下公式(3)的判断。

[0037] $$if\ \text{the count of}\ (y_i = 1) \geq (y_i = 0),\ y = 1;$$
$$else,\ y = 0 \quad \text{...公式（3）}$$

[0038] 因此，在BNN运算中，若n＝8，则比较逻辑430可将加总参数Ay与数值“4”作比较（大于、等于或小于），以决定输出参数y为“0”（原始取值-1）或“1”（原始取值1）。在此例中，若加总参数Ay大于或等于数值“4”，则表示子输出参数y_1、y_2～y_n当中的第一数值类型“1”的数量大于或等于第二数值类型“0”的数量。也就是说，由于子输出参数y_1、y_2～y_n的原始数值为“1”的数量大于或等于原始数值为“-1”的数量，因此子输出参数y_1、y_2～y_n求和所得的输出参数y为非负值，并且依据公式(3)取值“1”（原始取值1）。

[0039] 然而，在此例中，若加总参数Ay小于数值“4”，则表示子输出参数y_1、y_2～y_n当中的第一数值类型“1”的数量小于第二数值类型“0”的数量。也就是说，由于子输出参数y_1、y_2～y_n的原始数值为“1”的数量小于原始数值为“-1”的数量，因此对子输出参数y_1、y_2～y_n求和所得的输出参数y为负值，并且依据公式(3)取值“0”（原始取值-1）。比较逻辑230产生的输出参数y即为BNN运算的结果。

[0040] 据此，基于上述步骤S310～S330以及图2的微处理器电路200的架构，本实施例的微处理器电路200可采用较节省处理资源并且高效（本实施例可并行实现多个二值的乘法和累加运算）的方式来有效执行低精度的BNN运算。另外，关于上述表1、表2、公式(2)以及公式(3)的编码方式以及判断条件可依据不同运算需求来对应调整其内容，本发明并不限于此。

[0041] 图4是依照本发明的第二实施例的微处理器电路的示意图。图5是依照本发明的第二实施例的执行神经网络运算的方法的流程图。参考图4以及图5，微处理器电路400包括参数产生模块410、运算模块421、422以及比较逻辑430。参数产生模块410包括子参数产生模块411、412。在本实施例中，微处理器电路400适用于执行TNN运算，并且微处理器电路400欲实现的神经网络运算如同上述公式(1)。相较于上述实施例的微处理器电路200，不同之处在于本实施例的输入参数x_1、x_2～x_n、权重参数w_1、w_2～w_n及输出参数y的取值范围为{-1,0,1}。

[0042] 首先，在步骤S510中，微处理器电路400透过两个子参数产生模块411、412分别并行接收神经网络运算的多个输入参数x_1、x_2～x_n以及多个权重参数w_1、w_2～w_n。此两个子参数产生模块411、412依据输入参数x_1、x_2～x_n以及权重参数w_1、w_2～w_n来分别产生多个子输出参数$y_1[1]$、$y_2[1]$～$y_n[1]$、$y_1[0]$、$y_2[0]$～$y_n[0]$，其中n为大于1的正整数。在本实施例中，输入参数x_1、x_2～x_n以及权重参数w_1、w_2～w_n的取值集合为{-1,0,1}。在本实施例中，子参数产生模块411、412可分别先对输入参数x_1、x_2～x_n以及权重参数w_1、w_2～w_n进行编码，编码结果例如以下表3。并且，在本实施例中，编码后的输入参数x_1、x_2～x_n、编码后的权重参数w_1、w_2～w_n以及对应的子输出参数y_1、y_2～y_n的数值可如以下表4（第二查找表），其中$i \in \{1,2,3,\cdots,n\}$。

[0043]

w_i, x_i	w_i, x_i（编码后）
-1	1X
0	00
1	01

[0044] 表3

w_i（编码后）	x_i（编码后）	y_i
1X	1X	01
1X	00	00
1X	01	10
00	1X	00
00	00	00
00	01	00
01	1X	10
01	00	00
01	01	01

[0045]

[0046] 表4

[0047] 也就是说，在本实施例中，子参数产生模块411、412预先对输入参数x_1、x_2～x_n以及权重参数w_1、w_2～w_n进行编码，并且依据上述表4来快速地取得对应的子输出参数y_1、y_2～y_n。在本实施例中，子参数产生模块411输出查表所得的子输出参数y_1、y_2～y_n的第一位(例如高位)$y_1[1]$、$y_2[1]$～$y_n[1]$，以表示子输出参数y_1、y_2～y_n的数值为第一数值类型(“10”)的数量。子参数产生模块412输出查表所得的子输出参数y_1、y_2～y_n的第二位(例如低位)$y_1[0]$、$y_2[0]$～$y_n[0]$，以表示子输出参数y_1、y_2～y_n的数值为第二数值类型(“01”)的数量。

[0048] 另外，在一实施例中，子参数产生模块411也可依据编码后的输入参数x_1、x_2～x_n以

及编码后的权重参数w_1、w_2～w_n来透过以下公式(4)对应的数字逻辑电路(第一子逻辑电路),来取得对应的子输出参数y_1、y_2～y_n的高位$y_1[1]$、$y_2[1]$～$y_n[1]$。并且,子参数产生模块412也可依据编码后的输入参数x_1、x_2～x_n以及编码后的权重参数w_1、w_2～w_n来透过以下公式(5)对应的数字逻辑电路(第二子逻辑电路),来取得对应的子输出参数y_1、y_2～y_n的低位$y_1[0]$、$y_2[0]$～$y_n[0]$。如图4所示,参数产生模块411可包括第一子逻辑电路,并且参数产生模块412可包括第二子逻辑电路。在本实施例中,以下公式(4)以及公式(5)当中的符号“～”以及符号“&”为操作数的符号,分别代表“取反(INV)”以及“按位与(AND)”。

[0049] $y_i[1]=w_i[1]\&\sim x_i[1]\&x_i[0]+\sim w_i[1]\&w_i[0]\&x_i[1]$.....公式(4)

[0050] $y_i[0]=w_i[1]\&x_i[1]+\sim w_i[1]\&w_i[0]\&\sim x_i[1]\&x_i[0]$.....公式(5)

[0051] 接着,在步骤S520中,微处理器电路400透过子运算模块421、422分别并行接收对应的子输出参数的第一位$y_1[1]$、$y_2[1]$～$y_n[1]$以及第二位$y_1[0]$、$y_2[0]$～$y_n[0]$。子运算模块421、422分别加总子输出参数的第一位$y_1[1]$、$y_2[1]$～$y_n[1]$以及第二位$y_1[0]$、$y_2[0]$～$y_n[0]$,以产生两个加总参数Ay[1]、Ay[0]。在本实施例中,子运算模块421包括多个加法器421_1,并且子运算模块422包括多个加法器422_1。在本实施例中,子运算模块421、422分别对子输出参数$y_1[1]$、$y_2[1]$～$y_n[1]$、$y_1[0]$、$y_2[0]$～$y_n[0]$同时进行加法运算。举例来说,这些加法器421_1、422_1的第一层可对子输出参数$y_1[1]$、$y_2[1]$～$y_n[1]$、$y_1[0]$、$y_2[0]$～$y_n[0]$的每四个数据进行相加的运算,并且提供至这些加法器421_1、422_1的第二层。以此类推,这些加法器421_1、422_1的最后一层可输出两个加总参数Ay[1]、Ay[0]。在本实施例中,每个加法器421_1、422_1可由基本的4:2进位保留加法器(Carry Save Adder,CSA)实现,每个CSA加法器包括4个输入和2个输出(包括总和Sum和进位Carry)。

[0052] 在本实施例中,加总参数Ay[1]代表子输出参数y1[1]、$y_2[1]$～$y_n[1]$为第一数值类型(“10”)的数量,并且加总参数Ay[0]代表子输出参数$y_1[1]$、$y_2[1]$～$y_n[1]$为第二数值类型(“01”)的数量。然而,运算模块421、422的这些加法器421_1、422_1的数量及层数可依据子输出参数y_1、y_2～y_n的数量来对应设计,本发明的运算模块421、422的这些加法器421_1、422_1的配置方式不限于图4所示。

[0053] 在本实施例中,由于TNN算法中输入参数x_1、x_2～x_n以及权重参数w_1、w_2～w_n的取值范围为{-1,0,1},即取值仅三种数值“-1”、“0”和“1”,因此子参数产生模块411、412采用2-bit编码方式对输入参数x_1、x_2～x_n以及权重参数w_1、w_2～w_n进行编码后产生的编码后数值为“1X”、“00”、“01”。在一实施例中,数值“1X”也可代表“10”或“11”,本发明并不加以限制。换句话说,子运算模块421、422经由这些加法器421_1、422_1分别操作数输出参数的第一位$y_1[1]$、$y_2[1]$～$y_n[1]$以及第二位$y_1[0]$、$y_2[0]$～$y_n[0]$,以取得两个加总参数Ay[1]、Ay[0],其目的在于取得子输出参数y_1、y_2～y_n当中的第一数值类型(“10”)的数量以及第二数值类型(“01”)的数量。

[0054] 为了便于统计第一数值类型(“10”)以及第二数值类型(“01”)的数量,在一实施例中,子参数产生模块411、412输出的子输出参数y_1、y_2～y_n中,数值“01”代表“1”,数值“00”代表“0”,数值“10”代表“-1”。参考表4,子输出参数y_1、y_2～y_n的第一位$y_1[1]$、$y_2[1]$～$y_n[1]$为“1”的数量(即加总参数Ay[1])可表示子输出参数y_1、y_2～y_n的原始数值为“-1”的数量。子输出参数y_1、y_2～y_n的第二位$y_1[0]$、$y_2[0]$～$y_n[0]$为“1”的数量(即加总参数Ay[0])可表示子输出参数y_1、y_2～y_n的原始数值为“1”的数量。

[0055] 最后，在步骤S530中，微处理器电路400透过比较逻辑430接收两个加总参数Ay[1]、Ay[0]，并且微处理器电路400依据此两个加总参数Ay[1]、Ay[0]执行比较运算，以产生神经网络运算的输出参数y。在本实施例中，比较逻辑430比较此两个加总参数Ay[1]、Ay[0]，以判断子输出参数y_1、y_2～y_n当中的第一数值类型(“10”)的数量以及第二数值类型(“01”)的数量，并且决定输出参数y为“01”、“00”或“10”。举例来说，比较逻辑430可执行如以下公式(6)的判断。

[0056] if the count of(y_i[1]＝1)＞(y_i[0]＝1),y＝10；

[0057] if the count of(y_i[1]＝1)＜(y_i[0]＝1),y＝01；...公式(6)

[0058] else,y＝00

[0059] 因此，在TNN中，若子输出参数y_1、y_2～y_n当中的第一数值类型“10”(对应原始取值-1)的数量大于第二数值类型“01”(对应原始取值1)的数量，则对所有子输出参数y_1、y_2～y_n求和所得的输出参数y为负值，并且依据公式(6)取值“10”(对应原始取值-1)。若子输出参数y_1、y_2～y_n当中的第一数值类型“10”(对应原始取值-1)的数量小于第二数值类型“01”(对应原始取值1)的数量，则对所有子输出参数y_1、y_2～y_n求和所得的输出参数y为正值，并且依据公式(6)取值“01”(对应原始取值1)。否则，输出参数y取值“00”(对应原始取值0)。比较逻辑430产生的输出参数y即为上述公式(1)的TNN运算的结果。

[0060] 据此，基于上述步骤S510～S530以及图4的微处理器电路400的架构，本实施例的微处理器电路400可采用较节省处理资源并且高效(本实施例可并行实现多个三值的乘法和累加运算)的方式来有效执行低精度的TNN运算。另外，关于上述表3、表4、公式(4)、公式(5)以及公式(6)的编码方式以及判断条件可依据不同运算需求来对应调整其内容，本发明并不限于此。

[0061] 图6是依照本发明的第三实施例的微处理器电路的示意图。图7是依照本发明的第三实施例的执行神经网络运算的方法的流程图。参考图6以及图7，微处理器电路600包括参数产生模块610、运算模块620以及比较逻辑630。在本实施例中，微处理器电路600适用于执行BWN运算以及TWN运算，并且微处理器电路400欲实现的神经网络运算如同上述公式(1)。在BWN运算中，权重参数w_1、w_2～w_n的位宽为1～2bit，并且权重参数w_1、w_2～w_n的取值范围为{-1,1}。输入参数x_1、x_2～x_n与输出参数y的位宽相同。输入参数x_1、x_2～x_n与输出参数y皆为微处理器电路600所在处理器普通算术运算的全位宽(例如为8/16bit)。输入参数x_1、x_2～x_n与输出参数y的位宽大于权重参数w_1、w_2～w_n的位宽。需注意的是，TWN运算与BWN运算不同之处在于，权重参数w_1、w_2～w_n的取值范围为{-1,0,1}，并且微处理器电路600的位宽大于多个输入参数x_1、x_2～x_n及多个权重参数w_1、w_2～w_n的位宽总和。

[0062] 首先，在步骤S710中，微处理器电路600透过参数产生模块610并行接收神经网络运算的多个输入参数x_1、x_2～x_n以及多个权重参数w_1、w_2～w_n。参数产生模块610依据输入参数x_1、x_2～x_n以及权重参数w_1、w_2～w_n来产生多个子输出参数y_1’、y_2’～y_n’，其中n为大于0的正整数。在本实施例中，输入参数x_1、x_2～x_n为八位(8-bit)或十六位(16-bit)等诸如此类的全位(full-bit)参数。BWN运算的权重参数w_1、w_2～w_n的取值范围为{-1,1}。TWN运算的权重参数w_1、w_2～w_n的取值范围为{-1,0,1}。

[0063] 在本实施例中，参数产生模块610预先对权重参数w_1、w_2～w_n进行编码。参数产生模块610藉由依据权重参数w_1、w_2～w_n的取值范围来对权重参数w_1、w_2～w_n进行对应的编码。在

本实施例中，若权重参数w_1、w_2～w_n的取值范围为{-1,1}，则权重参数w_1、w_2～w_n的编码结果例如以下表5。其中i∈{1,2,3,…,n}。

[0064]

w_i	w_i（编码后）
-1	0
1	1

[0065] 表5

[0066] 在一实施例中，若权重参数w_1、w_2～w_n的取值范围为{-1,0,1}，则权重参数w_1、w_2～w_n的编码结果例如以下表6，其中i∈{1,2,3,…,n}。

[0067]

w_i	w_i（编码后）
-1	1X
0	00
1	01

[0068] 表6

[0069] 在本实施例中，参数产生模块610接着依据编码后的输入参数x_1、x_2～x_n及权重参数w_1、w_2～w_n来决定子输出参数y_1'、y_2'～y_n'。在本实施例中，若权重参数w_1、w_2～w_n的取值范围为{-1,1}，则参数产生模块610依据以下公式(7)(第四条件式)来产生对应的子输出参数y_1'、y_2'～y_n'。需先说明的是，以下公式(7)以及公式(8)当中的符号“[ ]”为操作数的符号，其中代表“补码”。

[0070]
$$\begin{aligned} &if\ w_i = 1, y_i' = x_i\ ; \\ &\quad else,\ y_i' = [x_i] \end{aligned} \quad \text{.....公式（7）}$$

[0071] 在一实施例中，若权重参数w_1、w_2～w_n的取值范围为{-1,0,1}，则参数产生模块610依据以下公式(8)(第五条件式)来产生对应的子输出参数y_1'、y_2'～y_n'。

[0072] if w_i=01,y_i'=x_i;

[0073] if w_i=1X,y_i'=[x_i];.....公式(8)

[0074] else,y_i'=00

[0075] 也就是说，当参数产生模块610判断权重参数w_1、w_2～w_n的取值范围包含两种数值类型时，参数产生模块610将子输入参数x_1、x_2～x_n各别的原始码或补码作为子输出参数y_1'、y_2'～y_n'。当参数产生模块610判断权重参数w_1、w_2～w_n的取值范围包含三种数值类型

时，参数产生模块610将零码或子输入参数x_1、x_2～x_n各别的原始码或补码作为多个子输出参数y_1'、y_2'～y_n'。

[0076] 接着，在步骤S720中，微处理器电路600透过运算模块620并行接收子输出参数y_1'、y_2'～y_n'，并且加总子输出参数y_1'、y_2'～y_n'，以产生加总参数Ay'。在本实施例中，运算模块620包括多个加法器621，并且运算模块620对子输出参数y_1'、y_2'～y_n'同时进行加法运算。举例来说，这些加法器621的第一层可对子输出参数y_1'、y_2'～y_n'的每四个数据进行相加的运算，并且提供至这些加法器621的第二层。以此类推，这些加法器621的最后一层可输出加总参数Ay'。在本实施例中，每个加法器621可由基本的4:2进位保留加法器(Carry Save Adder，CSA)实现，并且每个CSA加法器包括4个输入和2个输出(包括总和Sum和进位Carry)。然而，运算模块620的这些加法器621的数量及层数可依据子输出参数y_1'、y_2'～y_n'的数量来对应设计，本发明的运算模块620的这些加法器621的配置方式不限于图6所示。加总参数Ay'即对应上述公式(1)的乘加求和运算的结果。

[0077] 最后，在步骤S730中，微处理器电路600透过截位逻辑630接收加总参数Ay'，并且截位逻辑630对加总参数Ay'执行截位运算，以产生神经网络运算的输出参数y'。在本实施例中，截位逻辑630依据加总参数Ay'来进行近似运算，输出参数y'可等于或近似于公式(1)的输出参数y。换句话说，在精度允许的范围内，截位逻辑630可将加总参数Ay'当中的数值截位，以取得精度足够的计算结果。关于截位运算的方式具体而言包括：根据小数点的位置，对加总参数Ay'移位。对移位后的数值做饱和进位(saturation and round)处理，使得对普通数据宽度不会出现溢出的情况。当加总参数Ay'大于普通数据宽度的最大值或小于最小值时，加总参数Ay'会被饱和成普通数据宽度的最大值或最小值。

[0078] 因此，截位逻辑630产生的输出参数y'即为上述公式(1)的神经网络运算的结果(或近似结果)。据此，基于上述步骤S710～S730以及图6的微处理器电路600的架构，本实施例的微处理器电路600可采用较节省处理资源的方式来有效执行高精度的BWN运算以及TWN运算。另外，关于上述表5、表6、公式(7)以及公式(8)的编码方式以及判断条件可依据不同运算需求来决定，本发明并不限于此。

[0079] 综上所述，本发明的微处理器电路以及执行神经网络运算的方法透过参数产生模块对神经网络运算的输入数据以及权重数据的至少其中之一进行编码，并且依据编码后的数据产生对应的子输出参数。接着，本发明的微处理器电路以及执行神经网络运算可透过运算模块以及功能逻辑可计算此对应的子输出参数，以产生神经网络运算的输出参数。并且，本发明的功能逻辑的类型可依据不同的神经网络运算的类型来选择。因此，本发明的微处理器电路以及执行神经网络运算的方法，相较于一般的微处理器电路中的算术逻辑单元(ALU)以普通乘法器和加法器实现乘累加运算来产生神经网络运算的输出参数的技术方案相比，可有效节省神经网络运算的运算资源。

[0080] 虽然本发明已以实施例揭露如上，然其并非用以限定本发明，任何所属技术领域中具有通常知识者，在不脱离本发明的精神和范围内，当可作些许的更动与润饰，故本发明的保护范围当视权利要求所界定者为准。

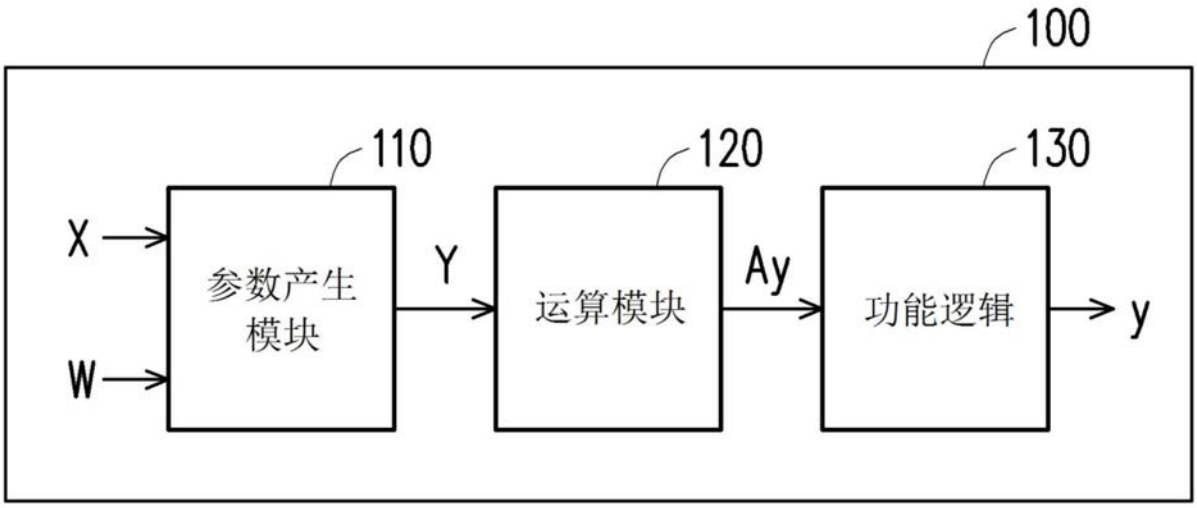


图1

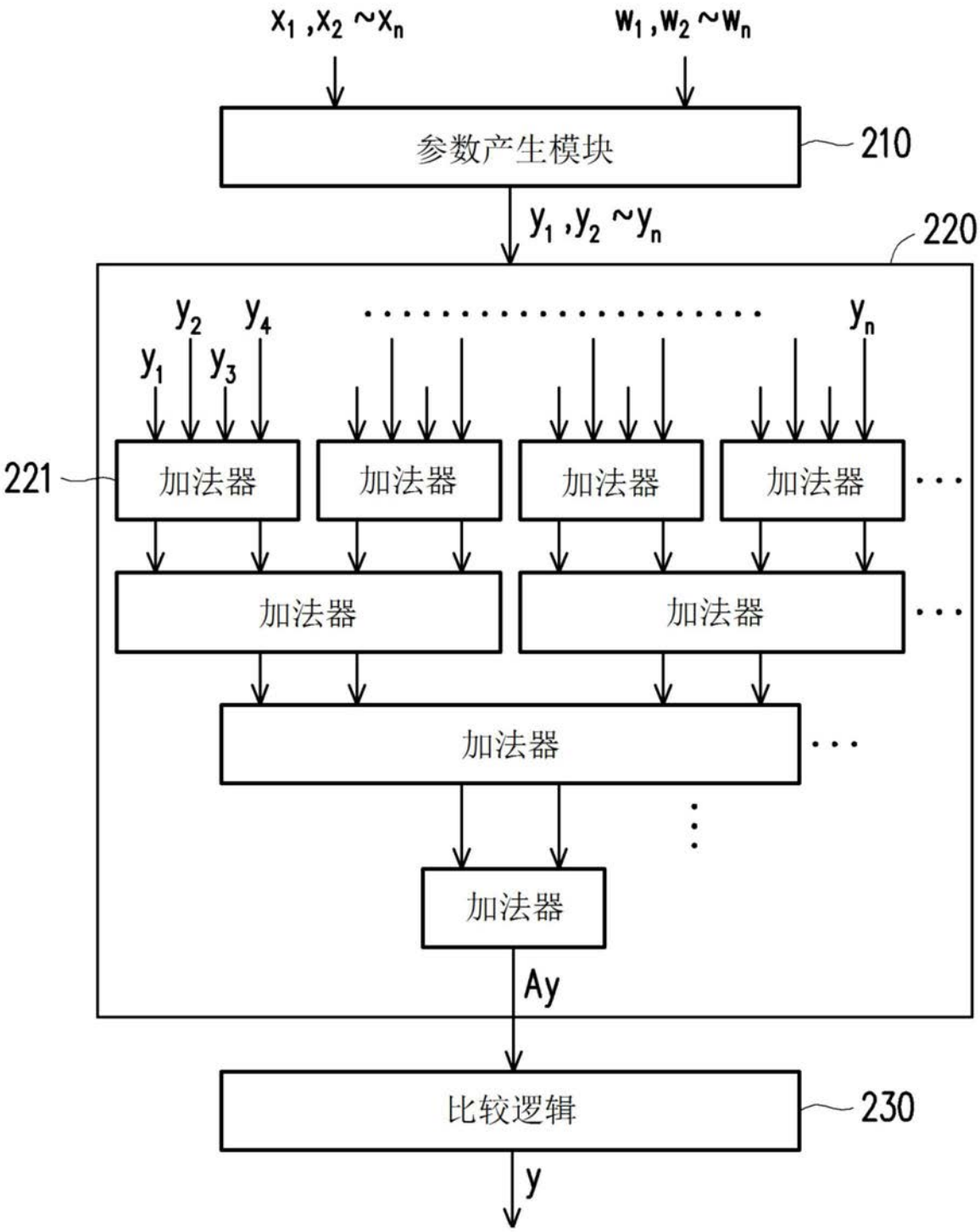

$x_1, x_2 \sim x_n$
$w_1, w_2 \sim w_n$
参数产生模块
210
$y_1, y_2 \sim y_n$
220
y_1 y_2 y_3 y_4 y_n
221
加法器
加法器
加法器
加法器
加法器
加法器
加法器
Ay
比较逻辑
230
y

图2

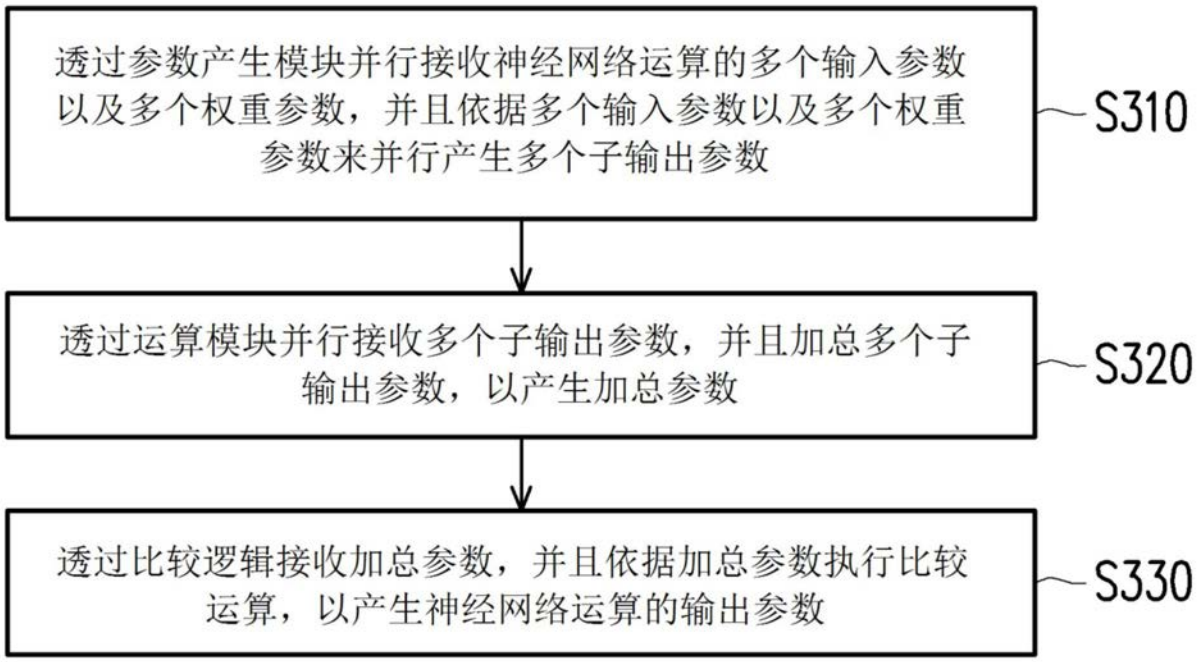

透过参数产生模块并行接收神经网络运算的多个输入参数以及多个权重参数，并且依据多个输入参数以及多个权重参数来并行产生多个子输出参数
S310
透过运算模块并行接收多个子输出参数，并且加总多个子输出参数，以产生加总参数
S320
透过比较逻辑接收加总参数，并且依据加总参数执行比较运算，以产生神经网络运算的输出参数
S330

图3

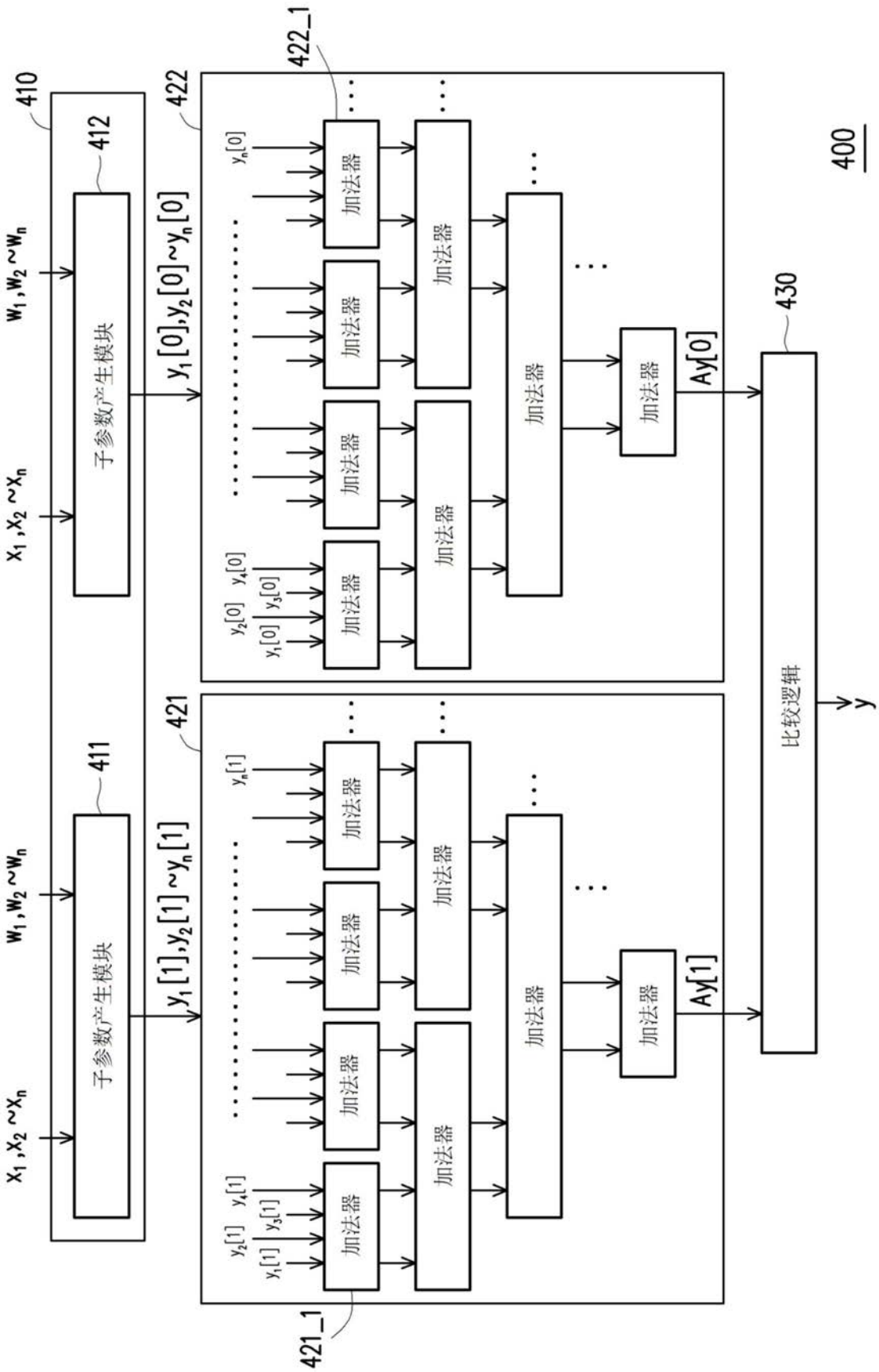

400
410
412
411
422
422_1
421
421_1
430
子参数产生模块
子参数产生模块
加法器
比较逻辑
$X_1, X_2 \sim X_n$
$W_1, W_2 \sim W_n$
$y_1[0], y_2[0] \sim y_n[0]$
$y_1[1], y_2[1] \sim y_n[1]$
Ay[0]
Ay[1]
y

图4

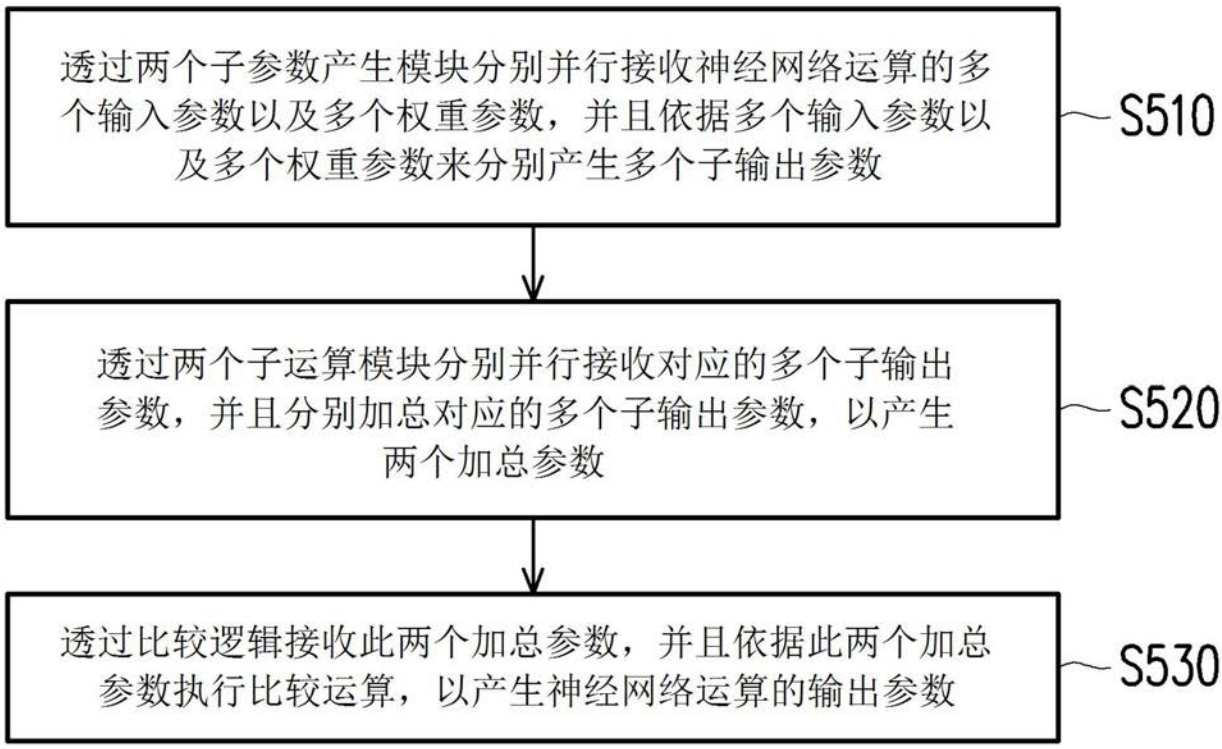

透过两个子参数产生模块分别并行接收神经网络运算的多个输入参数以及多个权重参数，并且依据多个输入参数以及多个权重参数来分别产生多个子输出参数
S510
透过两个子运算模块分别并行接收对应的多个子输出参数，并且分别加总对应的多个子输出参数，以产生两个加总参数
S520
透过比较逻辑接收此两个加总参数，并且依据此两个加总参数执行比较运算，以产生神经网络运算的输出参数
S530

图5

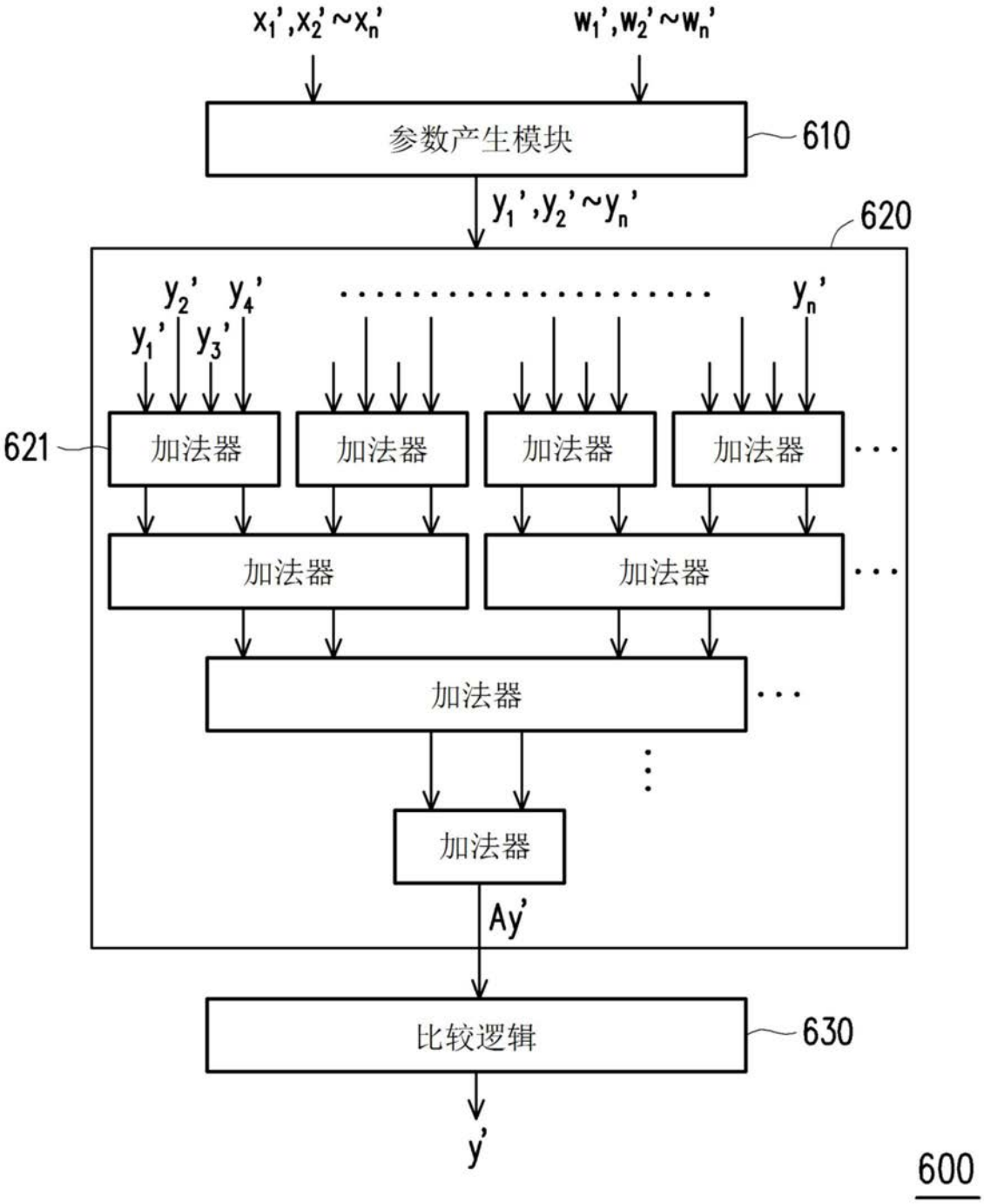

x1',x2'~xn'
w1',w2'~wn'
参数产生模块
610
y1',y2'~yn'
620
y2' y4'
y1' y3'
yn'
621
加法器
加法器
加法器
加法器
加法器
加法器
加法器
Ay'
比较逻辑
630
y'
600

图6

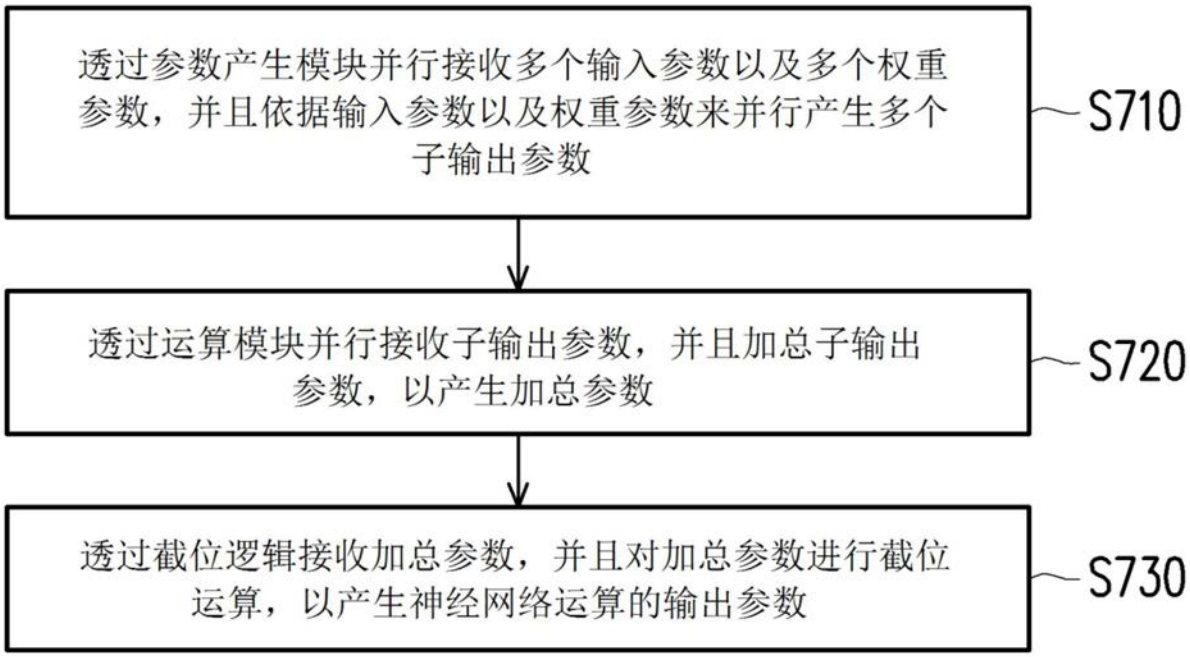

透过参数产生模块并行接收多个输入参数以及多个权重参数，并且依据输入参数以及权重参数来并行产生多个子输出参数
S710
透过运算模块并行接收子输出参数，并且加总子输出参数，以产生加总参数
S720
透过截位逻辑接收加总参数，并且对加总参数进行截位运算，以产生神经网络运算的输出参数
S730

图7