(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 114997392 A
(43)申请公布日 2022.09.02

(21)申请号 202210926707.X

(22)申请日 2022.08.03

(71)申请人 成都图影视讯科技有限公司
地址 610000 四川省成都市高新区高朋大道11号1栋3层K1号

(72)发明人 贺新

(74)专利代理机构 广州三环专利商标代理有限公司 44202
专利代理师 孙朝锐

(51)Int.Cl.
G06N 3/063(2006.01)
G06N 3/04(2006.01)

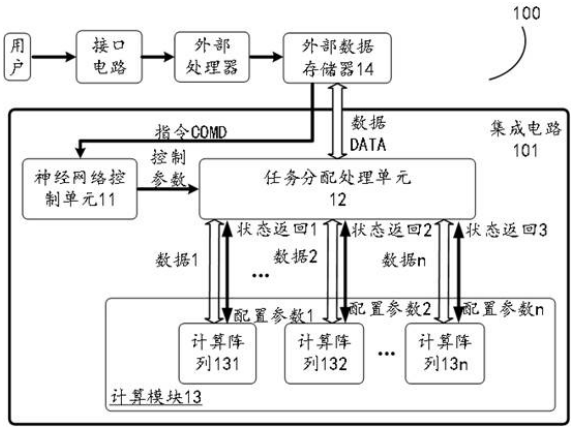
权利要求书2页 说明书6页 附图3页

(54)发明名称

用于神经网络计算的架构以及架构方法

(57)摘要

本申请公开了一种用于神经网络计算的架构以及架构方法，涉及人工智能技术领域。该神经网络架构方法包括：将每多个粒计算单元形成一个计算组；对所有计算组中的每个计算组根据计算功能需求进行参数集配置；将所有计算组中参数集配置相同的计算组设置为一个计算矩阵；每个计算矩阵根据对应的参数集从存储器中调取数据，并将该数据参数集中的相关参数进行运算，并将运算结果返回至存储器中。



CN 114997392 A

1.一种用于神经网络计算的架构,其特征在于,包括:

神经网络控制单元,用于从外部数据存储器中接收指令,并将所述指令解析,以产生代表指令具体执行行为的控制参数;

任务分配处理单元,用于接收所述控制参数,并对所述控制参数进行解析、拆解以及分组,以产生多组配置参数;以及用于接收外部数据存储器中的运算数据;以及

计算模块,耦接至所述任务分配处理单元,用于接收所述多组配置参数和所述运算数据,并将所述多组配置参数和所述运算数据进行运算处理,并将运算处理后的结果数据和状态报告返送至所述任务分配处理单元;其中,计算模块包括多个计算阵列,每个计算阵列接收所述多组配置参数中的一组。

2.如权利要求1所述的架构,其特征在于,每个所述计算阵列包括多个并行连接的计算组,每个计算组接收相同的配置参数。

3.如权利要求2所述的架构,其特征在于,每个计算组包括i×j个粒计算单元,其中,i为大于等于1的整数,j为大于等于1的整数,i×j为大于等于2的整数。

4.如权利要求3所述的架构,其特征在于,所述粒计算单元包括:

输入数据缓存,用于接收所述运算数据;

输入参数缓存,用于接收所述配置参数;

乘法器阵列,包括多个乘法器,每个乘法器用于从输入数据缓存和输入参数缓存中读取所述运算数据和所述配置参数,并做乘法运算;

累加器,用于对每个乘法器的计算结果做加法运算,并产生所述结果数据;以及

输出参数缓存,用于接收所述结果数据,并将结果数据返送至所述任务分配处理单元。

5.如权利要求3所述的架构,其特征在于,每个所述计算组包括32个粒计算单元。

6.如权利要求1所述的架构,其特征在于,所述任务分配处理单元包括:

解析模块,用于接收所述神经网络控制单元输出的控制参数,并对所述控制参数进行解析和拆解,以产生配置参数集;

功能分配模块,用于将所述配置参数集进行分组,以产生多组配置参数,并根据所述状态报告将每组配置参数发送至对应的计算阵列;以及

数据桥联模块,用于接收运算数据和结果数据。

7.一种用于神经网络计算的架构,其特征在于,包括:

接口电路,用于接收用户信息,并将用户信息处理;

外部处理器,用于接收接口电路传出的处理后的用户信号,以产生指令和数据;

外部存储器,用于接收来自所述外部处理器的所述指令和数据,并存储;以及

如权利要求1~6中任一项所述的架构。

8.一种用于神经网络计算的架构方法,其特征在于,包括:

将每i×j个粒计算单元形成一个计算组,其中,i为大于等于1的整数,j为大于等于1的整数,i×j为大于等于2的整数;

对m个计算组中的每个计算组根据计算功能需求进行参数集配置,其中,m为大于1的整数;

将m个计算组中参数集配置相同的计算组设置为一个计算矩阵;以及

每个计算矩阵根据对应的参数集中的第一参数从存储器中调取数据,并将该数据和对

应的参数集中第二参数进行运算，并将运算结果返回至存储器中。

9.如权利要求8所述的架构方法，其特征在于，每个粒计算单元包括9个乘法器。

10.如权利要求8所述的架构方法，其特征在于，i×j等于32。

用于神经网络计算的架构以及架构方法

技术领域

[0001] 本申请涉及人工智能技术领域,特别涉及一种用于神经网络计算的架构以及架构方法。

背景技术

[0002] 神经网络模型计算是通过模拟人脑的神经元处理机制来分析和处理如图像、声音、文本等数据信息,是人工智能(Artificial Intelligence,AI)领域重要的一部分。

[0003] 神经网络模型计算中的卷积神经网络(CNN,Convolutional Neural Network)算法因结构简单、适应性强、鲁棒性高等特点广泛应用于神经网络各个领域。但是由于卷积神经网络的数据计算的复杂性,如何对卷积神经网络的数据进行高速运算是业界的一个研究重点。对比现有的神经网络运用平台,通用的处理器,如中央处理单元(CPU,Central Processing Unit),图像处理器(GPU,Graphics Processing Unit),现场可编程阵列(FPGA,Field－Programmable Gate Array)等可通过指令灵活的控制整个数据计算过程,但很多需要一条指令控制一个神经单元进行运算,这就导致整个架构中指令数量庞大,数据传输效率低,计算速度慢;专用的神经网络处理器可高效复用数据因此效率很高,但在设计后又难以做出调整。

[0004] 因此,我们期望能提供一种能够高效且灵活的神经网络数据运算架构设计和方法。

发明内容

[0005] 针对现有技术中的一个或多个问题,本申请的实施例提供了一种用于神经网络计算的架构以及架构方法。

[0006] 本申请一方面提供了一种用于神经网络计算的架构。该架构包括:神经网络控制单元,用于从外部数据存储器中接收指令,并将所述指令解析,以产生代表指令具体执行行为的控制参数;任务分配处理单元,用于接收所述控制参数,并对所述控制参数进行解析、拆解以及分组,以产生多组配置参数;以及用于接收外部数据存储器中的运算数据;以及计算模块,耦接至所述任务分配处理单元,用于接收所述多组配置参数和所述运算数据,并将所述多组配置参数和所述运算数据进行运算处理,并将运算处理后的结果数据和状态报告返送至所述任务分配处理单元;其中,计算模块包括多个计算阵列,每个计算阵列接收所述多组配置参数中的一组。

[0007] 本申请另一方面提供了一种用于神经网络计算的架构,包括:接口电路,用于接收用户信息,并将用户信息处理;外部处理器,用于接收接口电路传出的处理后的用户信号,以产生指令和数据;外部存储器,用于接收来自所述外部处理器的所述指令和数据,并存储;以及如前所述的架构。

[0008] 本申请又一方面提供了一种用于神经网络计算的架构方法,包括:将每i×j个粒计算单元形成一个计算组,其中,i为大于等于1的整数,j为大于等于1的整数,i×j为大于

等于2的整数;对m个计算组中的每个计算组根据计算功能需求进行参数集配置,其中,m为大于1的整数;将m个计算组中参数集配置相同的计算组设置为一个计算矩阵;以及每个计算矩阵根据对应的参数集中的第一参数从存储器中调取数据,并将该数据和对应的参数集中第二参数进行运算,并将运算结果返回至存储器中。

[0009] 本申请公开的架构以及架构方法,其中,构架包括:神经网络控制单元,用于从外部数据存储器中接收指令,并将所述指令解析,以产生代表指令具体执行行为的控制参数;任务分配处理单元,用于接收所述控制参数,并对所述控制参数进行解析、拆解以及分组,以产生多组配置参数;以及用于接收外部数据存储器中的运算数据;以及计算模块,耦接至所述任务分配处理单元,用于接收所述多组配置参数和所述运算数据,并将所述多组配置参数和所述运算数据进行运算处理,并将运算处理后的结果数据和状态报告返送至所述任务分配处理单元;其中,计算模块包括多个计算阵列,每个计算阵列接收所述多组配置参数中的一组。由于该构架的应用时可减小指令数量,配置参数的数量也大大减小,因此在传输过程中寄存器中配置参数占位极少,数据占位更多。同时,该架构可批量动态分配计算组实现不同的计算功能,因此整个架构系统的数据计算效率大大提高。

附图说明

[0010] 图1所示为根据本申请一实施例的用于神经网络计算的神经网络系统架构框图;

图2所示为本申请一实施例提供的一个神经网络计算组的示意框图;

图3所示为本申请一实施例提供的用于做卷积计算的粒计算单元的示意框图;

图4所示为根据本申请一实施例提供的任务分配处理单元的示意框图;

图5所示为根据本申请一个实施例提供的一种高速神经网络计算方法;

其中,100-神经网络系统架构,12-任务分配处理单元。

具体实施方式

[0011] 为了使得本申请的目的技术方案和优点更加清楚,下面将结合附图对本申请的具体实施例进行详细描述。应当注意,这里描述的实施例只用于举例说明,并不用于限制本申请。在以下描述中,为了提供对本申请的透彻理解,阐述了大量特定细节。然而,对于本领域普通技术人员显而易见的是,不必采用这些特定细节来实行本申请。在其他实例中,为了避免混淆本申请,未具体描述公知的电路、材料或方法。

[0012] 在整个说明书中,对“一个实施例”、“实施例”、“一个示例”或“示例”的提及意味着:结合该实施例或示例描述的特定特征、结构或特性被包含在本申请至少一个实施例中。因此,在整个说明书的各个地方出现的短语“在一个实施例中”、“在实施例中”、“一个示例”或“示例”不一定都指同一实施例或示例。此外,可以以任何适当的组合和/或子组合将特定的特征、结构或特性组合在一个或多个实施例或示例中。此外,本领域普通技术人员应当理解,在此提供的附图都是为了说明的目的,并且附图不一定是按比例绘制的。应当理解,相同的附图标记指示相同的元件。这里使用的术语“和/或”包括一个或多个相关列出的项目的任何和所有组合。

[0013] 图1所示为根据本申请一实施例的用于神经网络计算的神经网络系统架构100。如图1所示,神经网络系统架构100包括神经网络控制单元11、任务分配处理单元12、计算模块

13和外部数据存储器14。

[0014] 在图1所示实施例中，神经网络控制单元11从外部数据存储器14中接收指令COMD。在一个实施例中，指令包括但不限于数据搬运指令(例如数据在外部存储器的地址、参数在外部存储器的地址等)、运算指令(例如数据读取量、计算数据量、矩阵运算中的向量和标量等)、控制指令(如神经网络计算单元需执行的功能等)以及逻辑操作指令等等。

[0015] 在一个实施例中，神经网络控制单元11解析指令，并将指令翻译成需要具体执行的行为，并产生代表具体执行行为的控制参数。这些控制参数包括但不限于神经网络预定义参数格式、神经网络每一层的参数等等。特别地，该参数涉及卷积层的卷积核的权值以及各通道的偏置量。

[0016] 在一个实施例中，任务分配处理单元12接收控制参数，并对控制参数进行解析、拆解及分组并产生计算模块13所适用的配置参数，并将配置参数发送至计算模块13。

[0017] 在一个实施例中，配置参数包括但不限于：1、定义特定的某个高效神经网络阵列的具体功能(对多输入待处理数据进行并行处理时)；2、神经网络的计算参数(如学习率、正则化参数、神经网络的层数、每一个隐层中神经元的个数，学习的回合数、小批量数据的大小、输出神经元的编码方式，代价函数的选择、权重初始化的方法、神经元激活函数的种类、参加训练模型数据的规模等超参数)、神经网络层的计算结果的存储器地址；3、神经网络层的输入通道输出通道数量、激活函数类型、滤波器大小等具体参数。

[0018] 同时，任务分配处理单元12还作为计算模块13和外部数据存储器14的桥联模块，接收外部数据存储器14中的运算数据DATA，以及接收计算模块13需读取或写入外部处理器的初始或中间计算的结果数据。在一个实施例中，运算数据和结果数据包括多种数据组，例如代表时间或频谱采样的一维数组，又或者是多通道的多维数组，例如代表平面上的二维像素点和RGB通道的三维输入数组。

[0019] 此外，任务分配处理单元12还将根据计算模块13需读取或写入外部处理器的初始或中间计算结果，使用硬件仲裁方式分析各个多通道高效神经网络计算阵列处于运行或空闲的状态，进行新任务分配。

[0020] 在一个实施例中，计算模块13根据配置参数的要求，通过任务分配处理单元12调取外部数据存储器14中的数据，并将数据和对应的配置参数进行运算，在一个计算周期结束后将计算结果返回至任务分配处理单元12，同时将状态报告发送至任务分配处理单元12。在一个实施例中，状态报告包括但不仅限于计算结束报告、准备好接收下条指令报告等等。

[0021] 在一个实施例中，计算模块13包括多通道并行的多个高效神经网络计算阵列131、132、…、13n。其中，n为大于等于1的整数，可根据应用场合进行选择适当的n值。每个高效神经网络计算阵列13n可执行相同和/或不同的计算功能。并根据执行的功能接收相应的配置参数。例如，在一些实施例中，高效神经网络计算阵列可执行矩阵计算(用于支持对神经网络模型中每个神经网络层的矩阵运算，包括：矩阵的加法、数乘、乘法、转置、卷积计算、反卷积计算等等)、数据预处理(对数据进行格式转换、滤波、划窗等操作)、算法处理(用于支持指定数字信号处理的算法以及自定义算法的运算，比如傅里叶变换、拉普拉斯变换、量化运算等等)、数据后处理(用于在神经网络模型计算完成之后进行数据后处理，包括输出结果转换、非线性运算等等)。

[0022] 在一个实施例中，每个高效神经网络计算阵列13n又包括多个并行运算的计算组。每个高效神经网络计算阵列13n中包括的计算组的数量可以相同也可以不同。在本公开的实施例中，每个计算组又包括相同数量的以矩阵排列的粒计算单元。任务分配处理单元12将根据神经网络控制单元11接收的指令调配合适数量的计算组形成一个高效神经网络计算阵列13n。每个高效神经网络计算阵列13n中的每个计算组将接收相同的参数配置，实现相同的运算规则。在一个实施例中，每个高效神经网络计算阵列13n中的每个计算组接收的数据可以相同也可以不同。也即是说：每个计算阵列13n根据所需完成的功能仅接收一组配置参数，每个计算阵列中的每个计算组(包括多个粒计算单元)均执行该配置参数。相比于传统的通用处理器(例如CPU)对每个粒计算单元单独配置参数进行数据运算，神经网络系统架构100将配置参数和数据传送分开，并且计算模块13中的每个高效神经网络计算阵列13n是以批量形式接受配置参数和数据，计算阵列13n自身仅按配置参数需要实现的功能对接收的数据进行计算即可。整个架构减小了指令数量，配置参数的数量也大大减小，因此在传输过程中寄存器中配置参数占位极少，数据占位更多。同时，该神经网络系统架构100可批量动态分配计算组实现不同的计算功能，实现同一计算功能的计算组形成一个计算阵列13n，因此大大提高了整个系统数据计算的效率。

[0023] 图2所示为本申请一实施例提供的一个神经网络计算组的示意框图。如图2所示，每个计算组包括i×j个粒计算单元。其中，i为大于等于1的整数，j为大于等于1的整数，i×j为大于等于2的整数，i和j的值可根据实际应用场合进行选择。在一个实施例中，每个计算阵列包括i×j=32个并行运算的粒计算单元。每个粒计算单元接收相同的配置参数，并根据该相同的配置参数对相同或不同数据进行计算。

[0024] 根据本申请的实施例，每个粒计算单元根据计算阵列可执行的计算功能具有多种实现结构。图3所示为本申请一实施例提供的用于做卷积计算的粒计算单元的示意框图。如图3所示，粒计算单元包括输入数据缓存、输入参数缓存、乘法器阵列、累加器和输出数据缓存。

[0025] 输入数据缓存接收任务分配处理单元12提供的数据。

[0026] 输入参数缓存接收任务分配处理单元12提供的配置参数。

[0027] 乘法器阵列中每个乘法器从输入数据缓存和输入参数缓存中读取数据和参数，并做乘法运算。

[0028] 累加器对乘法器的计算结果做加法运算，并将最终数据传送至输出参数缓存。输出参数缓存中的数据最终被送回至任务分配处理单元12。在其他不需要加法运算的实施例中，累加器也可以省略。

[0029] 在一个可选实施例中，粒计算单元包括9个乘法器。假设32个粒计算单元又构成一个计算组。由于计算组中的粒计算单元是统一进行编程的，配置参数的数量很少，即32个粒计算单元统一配备相同参数。如果使用传统处理器进行神经网络计算，则需调用32×9次指令去完成乘法计算，如果粒计算单元还有加法运算则调用的指令数量还需再乘以2，这些指令将会消耗大量的时间。本架构只需要给出一个指令告知配置参数，所有寄存单元将得到同一个参数，整个计算组在一个使用周期就可以将所有计算完成。

[0030] 此外，传统处理器处理神经网络计算时，每个乘法器在读取一个数据时都需要首先给存储器发出控制命令，等待储存器应答后传输数据，接收完了数据再给存储器反馈。这

样指令占位很多，比如，传送一个数据需要三个指令，那么数据传输效率只有25%。而现有架构指令和数据分开传送，同时批量处理指令，指令少了数据占位更多了，所以效率很高。

[0031] 每一个计算组中的粒计算单元可由神经网络控制单元11和任务分配处理单元12统一进行任务分配与重新编程。例如，当某个高效神经网络计算阵列13n中某个计算组的运算状态空闲时，任务分配处理单元12可对该计算组重新进行任务分配，此时该计算组可与其他相同运算任务的计算组组成新的计算阵列。

[0032] 图4所示为根据本申请一实施例提供的任务分配处理单元12的示意框图。如图4所示，任务分配处理单元12包括解析模块、功能分配模块和数据桥联模块。

[0033] 解析模块接收神经网络控制单元11输出的控制参数，并对控制参数进行解析和拆解，产生配置参数集。

[0034] 功能分配模块将解析模块解析后的配置参数集进行分组，同时接收计算模块13的状态报告和数据桥联模块的数据结果，产生多组配置参数，并将每组配置参数发送至计算模块13中对应的高效神经网络计算阵列13n以实现对应的运算。

[0035] 数据桥联模块接收外部数据存储器14中的数据DATA并将其发送至计算模块13中，以及接收计算模块13需读取或写入外部数据存储器14的初始或中间计算结果数据。

[0036] 继续返回参见图1。在一个实施例中，神经网络系统架构100中的神经网络控制单元11、任务分配处理单元12和计算模块被集成在集成电路101中。外部数据存储器14位于集成电路101外部。

[0037] 此外，根据实际应用的需要，在一些示例中，神经网络系统架构100还包括接口电路和外部处理器。

[0038] 在一个实施例中，接口电路接收用户信息，并将用户信息处理后传送至外部处理器。在一个实施例中，接口电路包括但不仅限于如USB、光模块、摄像头采集模块、以太网接口、蓝牙等等。

[0039] 外部处理器接收到接口电路传送的信号后，产生指令和数据，并将指令和数据存储在外部存储器14中。集成电路101用于将外部存储器14中的指令和数据做神经网络运算。

[0040] 在一个实施例中，接口电路、外部处理器、集成电路101以及外部存储器14将被塑封在同一个模块中。

[0041] 图5所示为根据本申请一个实施例提供的一种高速神经网络计算方法。该计算方法可用于前述神经网络架构100，包括以下步骤S1-S4。

[0042] 步骤S1：将每i×j个粒计算单元形成一个计算组，其中，i为大于等于1的整数，j为大于等于1的整数，i×j为大于等于2的整数，i和j的值可根据实际应用场合进行选择。在一个实施例中，i×j等于32。也即是说：每32个粒计算单元形成一个计算组。

[0043] 步骤S2：对m个计算组中的每个计算组根据计算功能需求进行参数集配置，其中，m为大于1的整数。

[0044] 步骤S3：将m个计算组中参数集配置相同的计算组设置为一个计算矩阵。

[0045] 步骤S4：每个计算矩阵根据对应的参数集中的第一参数从存储器中调取数据，并将该数据和对应的参数集中第二参数进行运算，并将运算结果返回至存储器中。在一个实施例中，第一参数包括例如寻址地址、调取数据量等参数。第二参数包括例如卷积层的卷积核的权值以及各通道的偏置量。

[0046] 虽然已参照几个典型实施例描述了本申请，但应当理解，所用的术语是说明和示例性、而非限制性的术语。由于本申请能够以多种形式具体实施而不脱离发明的精神或实质，所以应当理解，上述实施例不限于任何前述的细节，而应在随附权利要求所限定的精神和范围内广泛地解释，因此落入权利要求或其等效范围内的全部变化和改型都应为随附权利要求所涵盖。

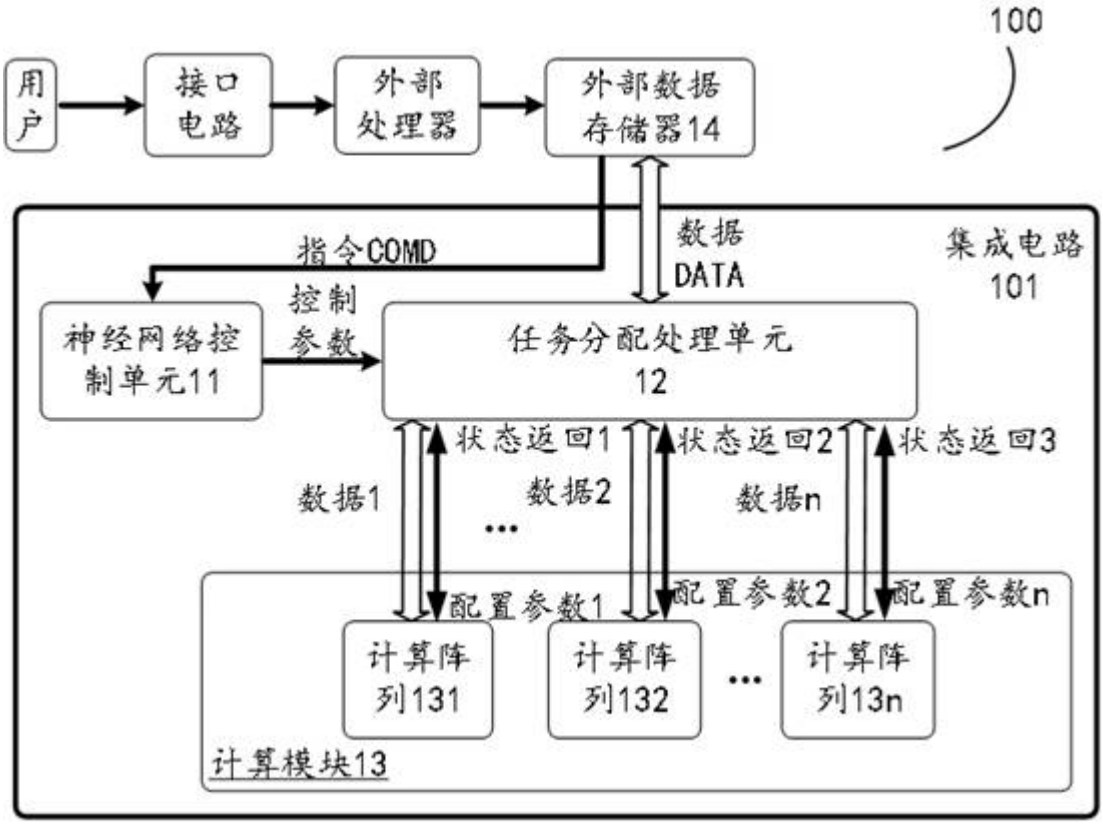

100
用户
接口电路
外部处理器
外部数据存储器14
集成电路101
指令COMD
数据DATA
控制参数
神经网络控制单元11
任务分配处理单元12
状态返回1
状态返回2
状态返回3
数据1
数据2
数据n
配置参数1
配置参数2
配置参数n
计算阵列131
计算阵列132
计算阵列13n
计算模块13

图1

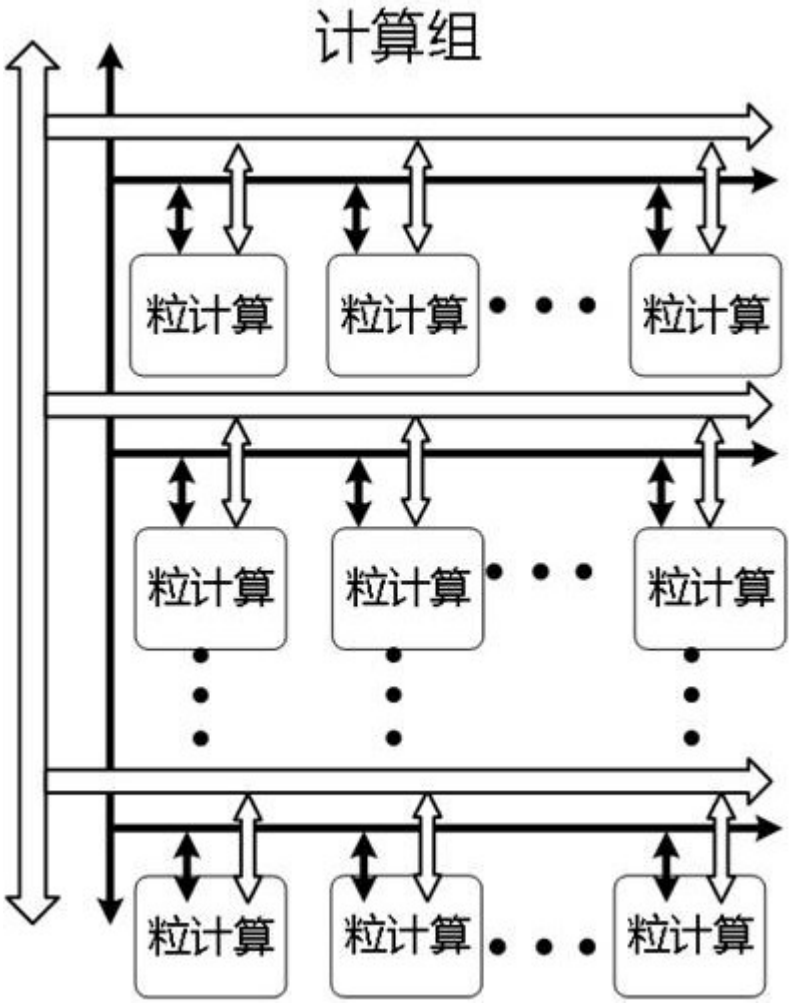

计算组
粒计算
粒计算
粒计算
粒计算
粒计算
粒计算
粒计算
粒计算
粒计算

图2

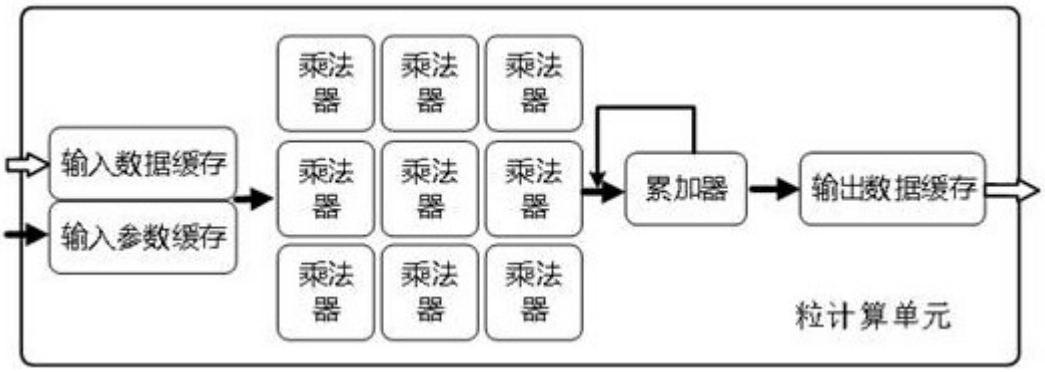

输入数据缓存
输入参数缓存
乘法器
乘法器
乘法器
乘法器
乘法器
乘法器
乘法器
乘法器
乘法器
累加器
输出数据缓存
粒计算单元

图3

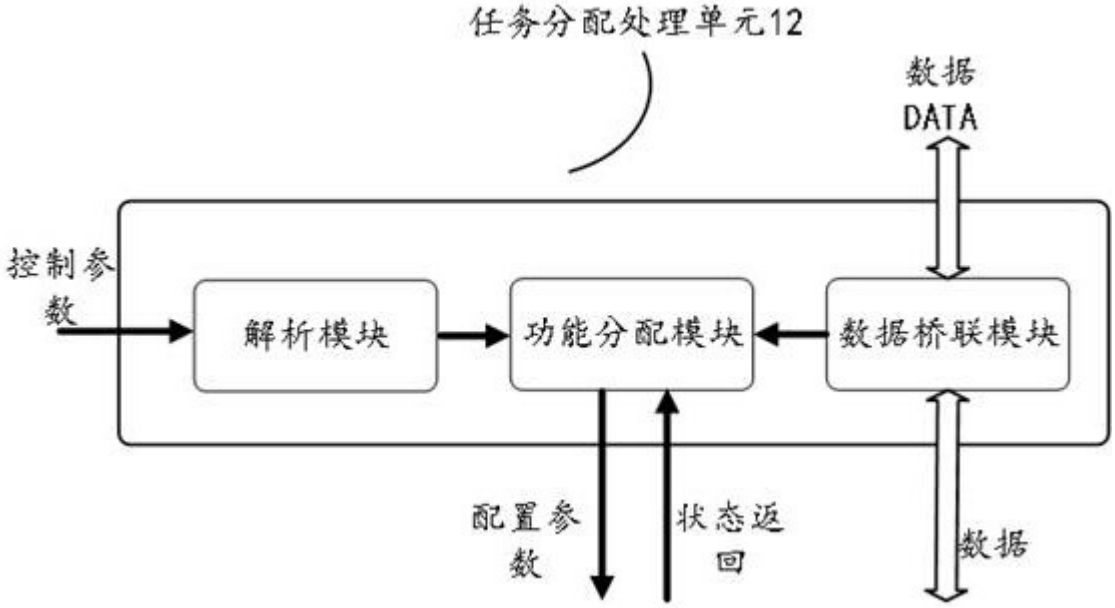

任务分配处理单元12
数据
DATA
控制参数
解析模块
功能分配模块
数据桥联模块
配置参数
状态返回
数据

图4

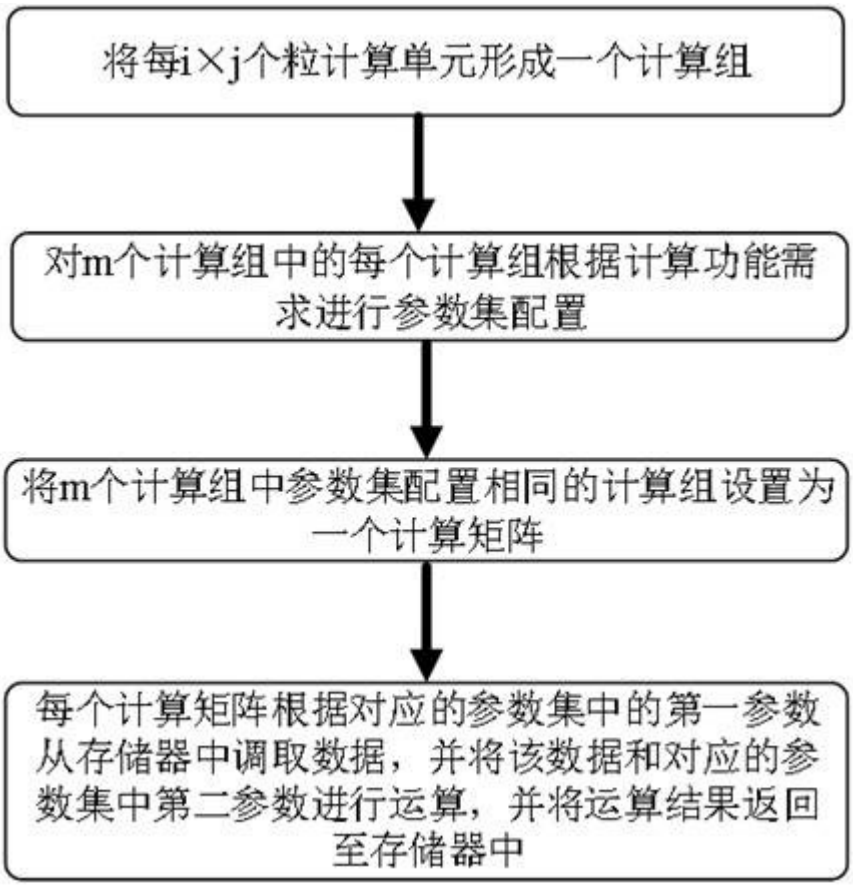

将每i×j个粒计算单元形成一个计算组
对m个计算组中的每个计算组根据计算功能需求进行参数集配置
将m个计算组中参数集配置相同的计算组设置为一个计算矩阵
每个计算矩阵根据对应的参数集中的第一参数从存储器中调取数据，并将该数据和对应的参数集中第二参数进行运算，并将运算结果返回至存储器中

图5