(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 111814973 A
(43)申请公布日 2020.10.23

(21)申请号 202010707798.9

(22)申请日 2020.07.18

(71)申请人 福州大学
地址 350108 福建省福州市闽侯县福州大学城乌龙江北大道2号福州大学

(72)发明人 魏榕山 张鼎盛 陈松林

(74)专利代理机构 福州元创专利商标代理有限公司 35100
代理人 陈明鑫 蔡学俊

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

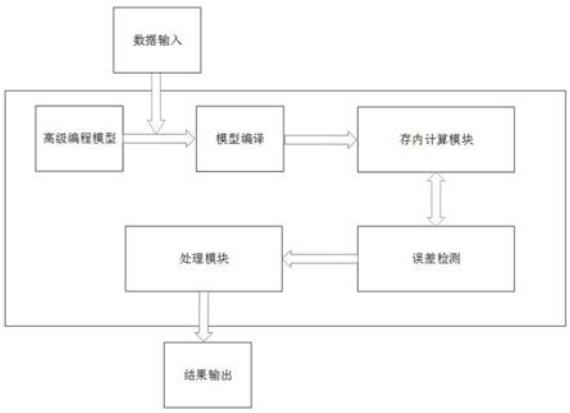
权利要求书1页 说明书5页 附图5页

(54)发明名称

一种适用于神经常微分方程网络计算的存内计算系统

(57)摘要

本发明涉及一种适用于神经常微分方程网络计算的存内计算系统，包括：高级编程模型，用于指定神经网络的架构以及指定神经常微分方程求解器；模型编译模块，确定输入数据的重组方案、权重映射方案和数据流控制原语；存内计算模块，将输入数据映射在存内计算模块的交叉阵列上，而后根据数据流控制原语进行常微分方程的计算；误差检验模块，用于计算并判断存内计算模块中输出结果的误差是否满足误差尤其；处理模块，将误差满足要求的输出结果根据任务需求进行处理，最后输出处理结果。本发明使用存内计算架构对神经常微分方程网络进行加速，极大降低了内存的占用率，提高了神经常微分方程网络的运算速度，精度和速度可调，可配置性强。



CN 111814973 A

1.一种适用于神经常微分方程网络计算的存内计算系统,其特征在于,包括:

高级编程模型,用于指定神经网络的架构以及指定神经常微分方程求解器;

模型编译模块,根据高级编程模型指定的神经网络的架构以及神经常微分方程求解器,确定输入数据的重组方案,以确定权重映射方案和数据流控制原语;

存内计算模块,根据权重映射方案将输入数据映射在存内计算模块的交叉阵列的每一个忆阻器单元上,而后根据数据流控制原语将输入数据从交叉阵列对应的字线输入,并进行常微分方程的计算;

误差检验模块,用于计算存内计算模块中输出结果的误差,然后与给定的误差值进行比较,若误差满足要求,则将存内计算模块中输出结果传输给处理模块;

处理模块,将误差满足要求的输出结果根据任务需求进行处理,最后输出处理结果。

2.根据权利要求1所述的一种适用于神经常微分方程网络计算的存内计算系统,其特征在于,所述指定神经网络的架构包括指定神经网络中卷积、池化、激活、卷积核大小的参数。

3.根据权利要求1所述的一种适用于神经常微分方程网络计算的存内计算系统,其特征在于,所述神经常微分方程求解器包括四种常微分方程解法:欧拉法、梯形法、四阶龙格库塔法、亚当斯法。

4.根据权利要求1所述的一种适用于神经常微分方程网络计算的存内计算系统,其特征在于,所述模型编译模块的具体实现方式如下:

S41、接收高级编程模型指定的神经网络的架构、神经常微分方程求解器和输入数据;

S42、判断输入数据是否符合规定尺寸大小,若不是则对其进行下采样和填充,使其符合规定尺寸大小,即使得得到的输入数据的特征图大小C×H×W,卷积次数N,卷积核大小K×K,0填充大小P;

S43、根据预设的卷积次数将特征图的通道数进行增加,使特征图大小变为(C+N)×H×W;

S44、找出值全为0的通道数i,重新调整特征图大小为(C+N-i)×H×W;

S45、将填充0后的特征图数据(C+N-i)×(H+2P)×(W+2P)按照结构化进行重组,分解为(K-P)×(K-P),(K-P)×K,K×K,K×(K-P)四种数据格式;

S46、将卷积滤波器按照步骤S45的结果分解为对应的模式,形成映射方案;

S47、根据指定的神经常微分方程求解器设计数据流;

S48、针对S46、S47处理后的数据生成权重映射方案和数据流控制原语。

一种适用于神经常微分方程网络计算的存内计算系统

技术领域

[0001] 本发明涉及一种适用于神经常微分方程网络计算的存内计算系统。

背景技术

[0002] 随着深度学习网络在人工智能领域取得了极大的成功,人们日常生活中的数据量越来越大,深度学习网络的规模也随之变大,参数量越来越多。数据在处理单元和内存之间不断搬运成为了计算系统中的一个关键性能瓶颈。考虑上述问题,存内计算(Processing-in-Memory)展现出了它在深度学习加速领域中的巨大前景。

[0003] 目前,使用存内计算加速的神经网络很少,主要是二值卷积神经网络,二值神经网络能够最大程度地降低模型的存储占用和模型的计算量,将神经网络中原本32位浮点数参数量化至1位定点数,同时极大加速了神经网络的推断过程。但是二值化不可避免地导致严重的信息损失,其量化函数不连续性也给深度网络的优化带来了困难。现如今还没有人为神经常微分方程网络进行存内计算加速,但该网络能够在保存信息的同时降低模型部署的资源消耗,是非常适合使用存内计算加速的网络。

发明内容

[0004] 本发明的目的在于提供一种适用于神经常微分方程网络计算的存内计算系统,使用存内计算架构对神经常微分方程网络进行加速,极大降低了内存的占用率,提高了神经常微分方程网络的运算速度,精度和速度可调,可配置性强。

[0005] 为实现上述目的,本发明的技术方案是:一种适用于神经常微分方程网络计算的存内计算系统,包括:

[0006] 高级编程模型,用于指定神经网络的架构以及指定神经常微分方程求解器;

[0007] 模型编译模块,根据高级编程模型指定的神经网络的架构以及神经常微分方程求解器,确定输入数据的重组方案,以确定权重映射方案和数据流控制原语;

[0008] 存内计算模块,根据权重映射方案将输入数据映射在存内计算模块中交叉阵列的每一个忆阻器单元上,而后根据数据流控制原语将输入数据从交叉阵列对应的字线输入,并进行神经常微分方程网络的计算;

[0009] 误差检验模块,用于计算存内计算模块中输出结果的误差,然后与给定的误差值进行比较,若误差满足要求,则将存内计算模块中输出结果传输给处理模块;

[0010] 处理模块,将误差满足要求的输出结果根据任务需求进行处理,最后输出处理结果。

[0011] 在本发明一实施例中,所述指定神经网络的架构包括指定神经网络中卷积、池化、激活、卷积核大小的参数。

[0012] 在本发明一实施例中,所述神经常微分方程求解器包括四种常微分方程解法:欧拉法、梯形法、四阶龙格库塔法、亚当斯法。

[0013] 在本发明一实施例中,所述模型编译模块的具体实现方式如下:

[0014] S41、接收高级编程模型指定的神经网络的架构、神经常微分方程求解器和输入数据；

[0015] S42、判断输入数据是否符合规定尺寸大小，若不是则对其进行下采样和填充，使其符合规定尺寸大小，即使得得到的输入数据的特征图大小C×H×W，卷积次数N，卷积核大小K×K，0填充大小P；

[0016] S43、根据预设的卷积次数将特征图的通道数进行增加，使特征图大小变为(C+N)×H×W；

[0017] S44、找出值全为0的通道数i，重新调整特征图大小为(C+N-i)×H×W；

[0018] S45、将填充0后的特征图数据的大小(C+N-i)×(H+2P)×(W+2P)按照结构化进行重组，分解为(K-P)×(K-P)，(K-P)×K，K×K，K×(K-P)四种数据格式；

[0019] S46、将卷积滤波器按照步骤S45的结果分解为对应的模式，形成映射方案；

[0020] S47、根据指定的神经常微分方程求解器设计数据流；

[0021] S48、针对S46、S47处理后的数据生成权重映射方案和数据流控制原语。

[0022] 相较于现有技术，本发明具有以下有益效果：本发明使用存内计算架构对神经常微分方程网络进行加速，极大降低了内存的占用率，提高了神经常微分方程网络的运算速度，精度和速度可调，并且提供了多种常微分方程的求解方案，可配置性强，解决了在硬件资源有限的情况下无法实现较大网络模型和模型可配置性差的问题。

附图说明

[0023] 图1给出了存内计算架构框图。

[0024] 图2给出了编译模块的流程图。

[0025] 图3给出了存内计算模块的结构框图。

[0026] 图4给出了交叉阵列的结构框图

[0027] 图5给出了存内计算架构工作流程图。

具体实施方式

[0028] 下面结合附图，对本发明的技术方案进行具体说明。

[0029] 本发明提供了一种适用于神经常微分方程网络计算的存内计算系统，包括：

[0030] 高级编程模型，用于指定神经网络的架构以及指定神经常微分方程求解器；

[0031] 模型编译模块，根据高级编程模型指定的神经网络的架构以及神经常微分方程求解器，确定输入数据的重组方案，以确定权重映射方案和数据流控制原语；

[0032] 存内计算模块，根据权重映射方案将输入数据映射在存内计算模块中交叉阵列的每一个忆阻器单元上，而后根据数据流控制原语将输入数据从交叉阵列对应的字线输入，并进行神经常微分方程网络的计算；

[0033] 误差检验模块，用于计算存内计算模块中输出结果的误差，然后与给定的误差值进行比较，若误差满足要求，则将存内计算模块中输出结果传输给处理模块；

[0034] 处理模块，将误差满足要求的输出结果根据任务需求进行处理，最后输出处理结果。

[0035] 所述指定神经网络的架构包括指定神经网络中卷积、池化、激活、卷积核大小的参

数。所述神经常微分方程求解器包括四种常微分方程解法:欧拉法、梯形法、四阶龙格库塔法、亚当斯法。

[0036] 所述模型编译模块的具体实现方式如下:

[0037] S41、接收高级编程模型指定的神经网络的架构、神经常微分方程求解器和输入数据;

[0038] S42、判断输入数据是否符合规定尺寸大小,若不是则对其进行下采样和填充,使其符合规定尺寸大小,即使得得到的输入数据的特征图大小C×H×W,卷积次数N,卷积核大小K×K,0填充大小P;

[0039] S43、根据预设的卷积次数将特征图的通道数进行增加,使特征图大小变为(C+N)×H×W;

[0040] S44、找出值全为0的通道数i,重新调整特征图大小为(C+N-i)×H×W;

[0041] S45、将填充0后的特征图数据的大小(C+N-i)×(H+2P)×(W+2P)按照结构化进行重组,分解为(K-P)×(K-P),(K-P)×K,K×K,K×(K-P)四种数据格式;

[0042] S46、将卷积滤波器按照步骤S45的结果分解为对应的模式,形成映射方案;

[0043] S47、根据指定的神经常微分方程求解器设计数据流;

[0044] S48、针对S46、S47处理后的数据生成权重映射方案和数据流控制原语。

[0045] 以下为本发明的具体实现过程。

[0046] 本发明应用在存内计算领域,具体为一种适用于神经常微分方程网络的存内计算系统。在存内计算架构上实现神经常微分方程网络,减少了神经网络模型的存储占用量,减少了实际芯片的面积,解决了大型网络无法在资源有限的边缘端部署,硬件功耗大,网络可配置性差的问题。

[0047] 本发明的系统框架如图1所述。各部分功能如下所诉:

[0048] 1、高级编程模型

[0049] 高级编程模型用于指定神经网络的架构包括网络中的卷积、池化、激活、卷积核大小等参数。指定合适的神经常微分方程求解器,本次发明提供四种常微分方程解法欧拉法,梯形法,四阶龙格库塔法,亚当斯法。指定初始误差大小。

[0050] 2、模型编译模块

[0051] 模型编译模块将处理高级编程模型和输入数据,根据神经常微分方程网络算法确定输入数据的重组方案,以确定数据重组、卷积核映射方案和数据流。具体流程如图2所示。

[0052] 1)接收高级编程模型和输入数据。

[0053] 2)判断此时的输入数据是否符合规定尺寸大小如果不是则对其进行下采样和填充。

[0054] 3)得到的特征图大小为C×H×W,卷积次数N,卷积核大小K×K,0填充大小P。

[0055] 4)根据卷积次数将特征图的通道数进行增加,使特征图大小变为(C+N)×H×W。

[0056] 5)找出值全为0的通道数i,重新调整特征图大小为(C+N-i)×H×W。

[0057] 6)将填充0后的特征图数据的大小(C+N-i)×(H+2P)×(W+2P)按照结构化进行重组,分解为(K-P)×(K-P),(K-P)×K,K×K,K×(K-P)四种数据格式。

[0058] 7)将卷积滤波器按照上一步的结果分解为对应的模式,形成映射方案。

[0059] 8)根据选择的常微分方程求解器设计数据流。

[0060] 9)针对处理后的数据生成权重映射方案和数据流控制原语

[0061] 3、存内计算模块

[0062] 存内计算模块是架构中的核心部分，用来计算神经网络中的矩阵与矩阵相乘，因为不需要将数据从内存和存储器之间来回搬运所以极大的加速了矩阵相乘操作并且降低了功耗开销。

[0063] 存内计算模块的具体结构如图3所示：

[0064] (1)缓存。用来存储运算过程中的临时变量和输入数据

[0065] (2)非线性函数单元。它实现了神经常微分方程网络中使用的非线性函数包括激活函数ReLU和池化函数Maxpooling。因为它们分别是使用最广泛的激活和池化函数。我们将激活单元实现为一个查找表。在某些场景下，例如当一个大的矩阵被映射到多个交叉阵列中时，可以绕过激活单元。

[0066] (3)输入输出总线。用来实现各模块之间的通信，实现数据传输。

[0067] (4)交叉阵列中的结构如图4所示：

[0068] 移位相加单元。对于一个大的矩阵，如果一个交叉阵列无法容乃，则应该将输入和输出进行分割，并且分组为多个交叉阵列。每个交叉阵列的输出为部分相加，通过移位加法单元进行水平收集和垂直相加，得出实际结果。

[0069] 采样保持电路。它捕捉位线电流，将电流转换为电压，并将电压发送到模数转换器单元，由模数转换器转换为数字结果。

[0070] 数模转换器。它将数字输入转换为应用于每个字线的相应电压。在这项工作中，我们假设每个字线在每个周期内都会收到一个比特的输入电压。

[0071] 模数转换器。将忆阻器交叉阵列的模拟信号转换为数字输出结果。在矩阵向量乘法中，它的功耗很高。在这项工作中，将8个忆阻器交叉杆共享一个模数转换器，以分摊模数转换器的开销。

[0072] 4、误差检验模块

[0073] 将存内计算模块中输出的结果计算误差，然后与选择器中给定的误差值进行比较，其误差可以使用步长折半前后两次计算结果的偏差Δ来判断所选步长是否适当，根据比较的结果判断能否进行下一步操作，当要求数指精度为ε时，如果Δ＞ε，反复将步长折半进行计算，直至△＜ε为止。表面上看，为了选择步长，每一步都要反复判断Δ，增加了计算工作量，但是在常微分方程的解变化剧烈的情况下，总的计算工作量会减少。

[0074] 5、处理模块

[0075] 误差满足要求的计算结果进入处理模块，根据任务要求进行处理，其中有全连接层，在整个网络中起到“分类器”的作用，将从神经网络中学习到的特征表示映射到标记空间，在实际操作中可以由卷积操作实现，对前层是卷积层的全连接层可以转化为全局卷积，其本质就是一个特征空间线性变换到另一个特征空间，最后的结果可以作为输出。

[0076] 本发明基于内存计算设计了神经常微分方程网络的实现系统。该系统具有可配置性，速度可调，功耗低等优点。适用于各种人工智能任务。内存占用量少，能够在资源有限的硬件条件下实现更大的神经网络。使用选择器可以选择各种常微分方程的求解器，并指定处理任务的速度和精度。模型编译模块可以针对神经常微分方程网络的特点优化数据控制数据流，进一步加速网络。使用存内计算模块能够极大加速神经常微分方程网络，降低功

耗,减少面积损耗。

[0077] 本发明一种适用于神经常微分方程网络计算的存内计算系统的工作流程图如图5所示,包括以下步骤:

[0078] 1)读取特征图信息,神经常微分方程网络信息,和速度要求;

[0079] 2)根据网络信息,选择最合适的常微分方程求解器;

[0080] 3)调整特征图大小到指定值;

[0081] 4)根据网络中需要的卷积次数增加特征图的通道数;

[0082] 5)对填充0后的特征图数据进行分割与重组,得到多种数据格式;

[0083] 6)根据上一步重组的信息,分解卷积核,形成多种映射方案;

[0084] 7)根据选择的求解器生成数据流原语;

[0085] 8)根据生成的映射方案将卷积核中的权重数据映射在存内计算模块中忆阻器交叉阵列的每一个忆阻器单元上;

[0086] 9)根据数据流原语将输入特征图从忆阻器阵列对应的字线输入,通过存内计算模块完成一次常微分方程的计算;

[0087] 10)将计算的结果送入误差检验模块,根据原先设定的误差进行判断,如果误差小于给定值,即得到输出结果。如果误差大于给定值,则重复步骤8);

[0088] 11)将误差在给定范围内的结果送入处理模块,经过处理模块后就能得到输出结果。

[0089] 以上是本发明的较佳实施例,凡依本发明技术方案所作的改变,所产生的功能作用未超出本发明技术方案的范围时,均属于本发明的保护范围。

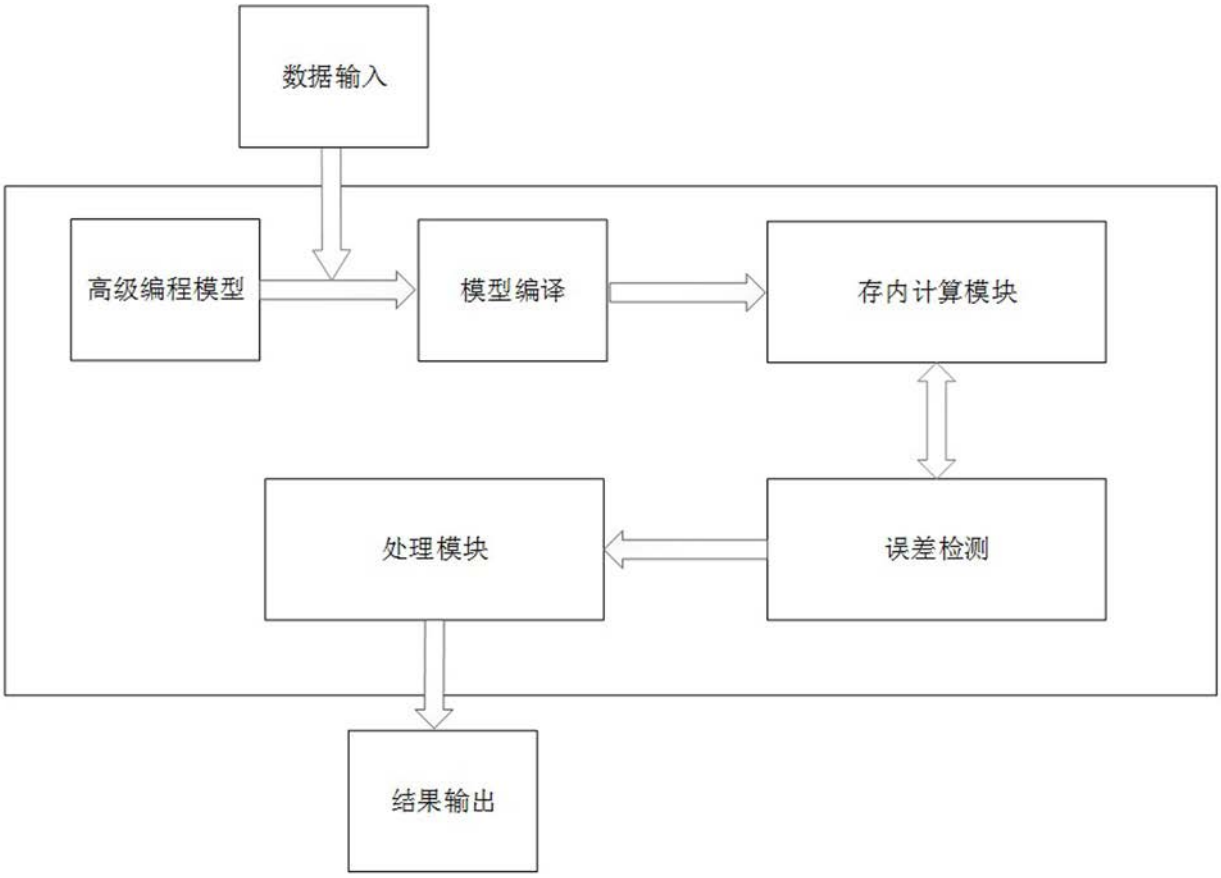


图1

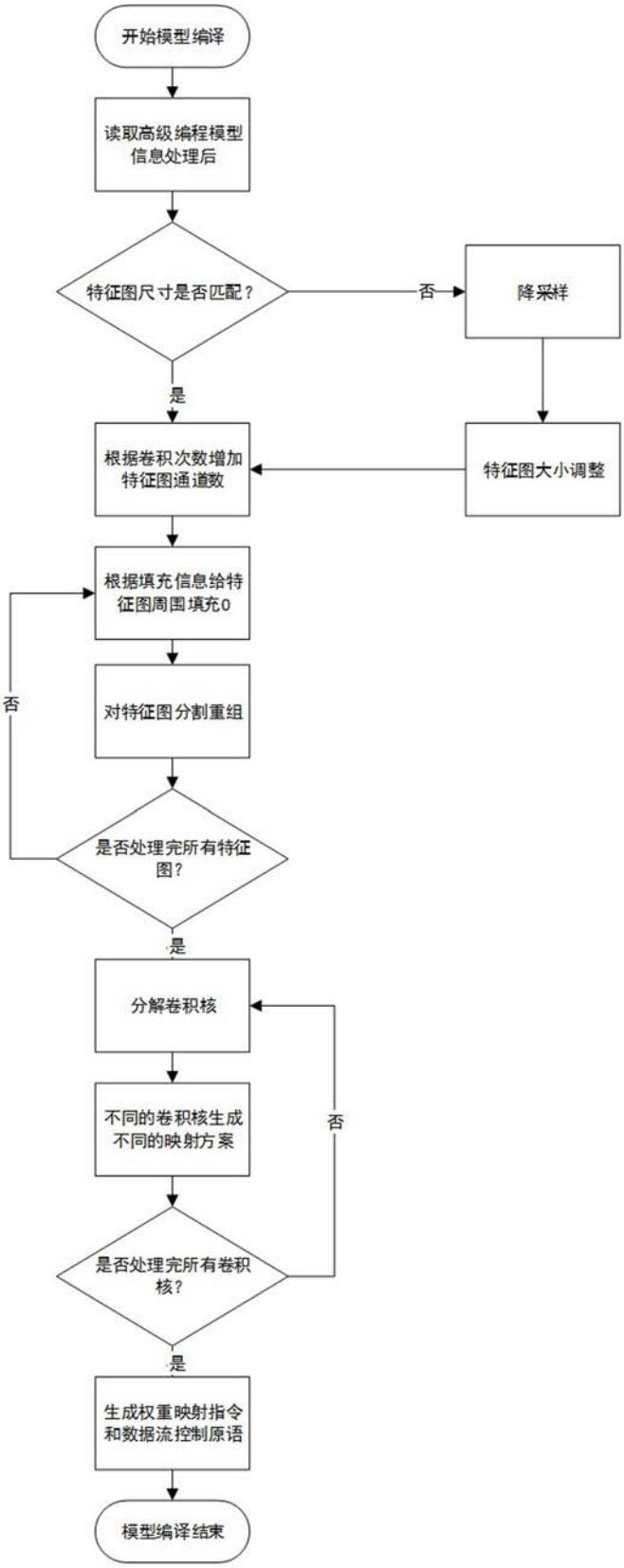

开始模型编译
读取高级编程模型信息处理后
特征图尺寸是否匹配？
否
降采样
是
根据卷积次数增加特征图通道数
特征图大小调整
根据填充信息给特征图周围填充0
对特征图分割重组
否
是否处理完所有特征图？
是
分解卷积核
不同的卷积核生成不同的映射方案
否
是否处理完所有卷积核？
是
生成权重映射指令和数据流控制原语
模型编译结束

图2

图3

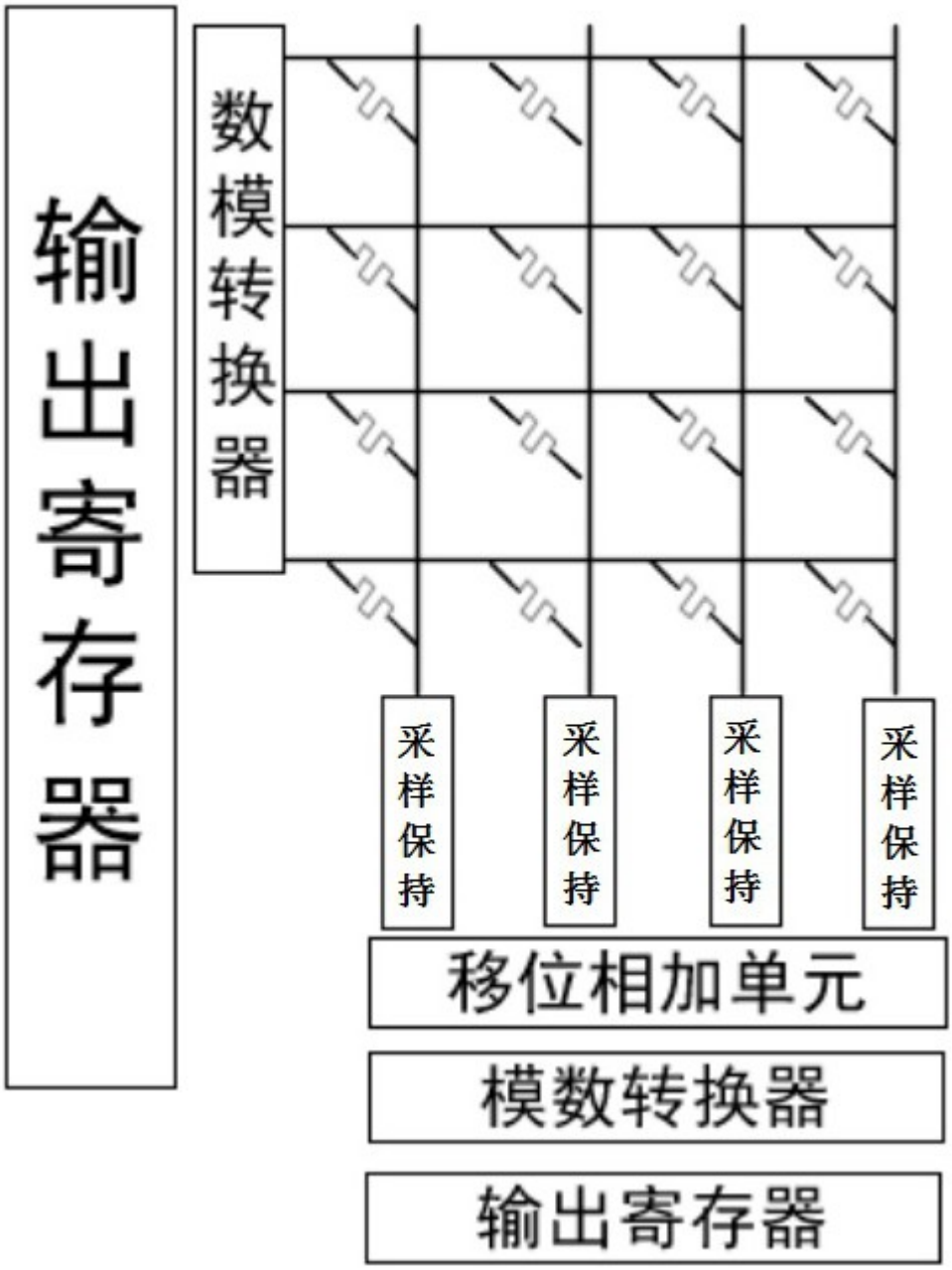


图4

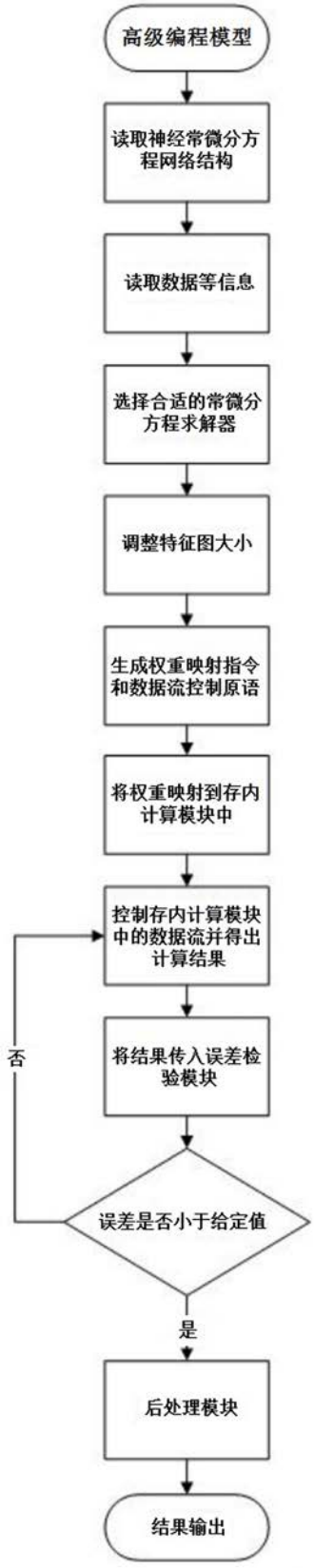

高级编程模型
读取神经常微分方程网络结构
读取数据等信息
选择合适的常微分方程求解器
调整特征图大小
生成权重映射指令和数据流控制原语
将权重映射到存内计算模块中
控制存内计算模块中的数据流并得出计算结果
将结果传入误差检验模块
误差是否小于给定值
否
是
后处理模块
结果输出

图5