(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 114186598 A
(43)申请公布日 2022.03.15

(21)申请号 202110856642.1

(22)申请日 2021.07.28

(71)申请人 中国科学院计算技术研究所
地址 100190 北京市海淀区中关村科学院南路6号

(72)发明人 何银涛 王颖 李华伟 李晓维

(74)专利代理机构 北京泛华伟业知识产权代理有限公司 11280
代理人 王勇

(51)Int.Cl.
G06K 9/62(2022.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)

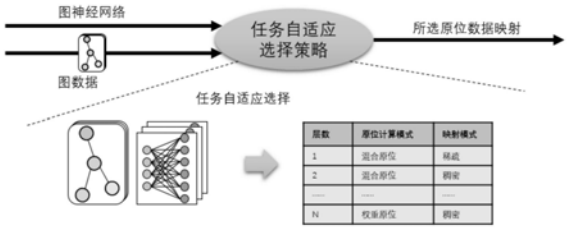
权利要求书2页 说明书10页 附图2页

(54)发明名称

一种基于阻变存储器的图神经网络计算方法和装置

(57)摘要

本发明实施例提供了一种基于阻变存储器的图神经网络计算方法和装置，该方法包括：对于图神经网络的任一层，分析该层中将要在阻变存储器阵列中运算的图数据在权重原位计算模式和混合原位计算模式下的处理时延相对大小，选择时延最小的模式作为该层的计算模式；在权重原位计算模式，对所述图神经网络的所述层将图数据的邻接矩阵和图神经网络的权重参数作为原位数据分别映射到相应的阻变存储器阵列中，以将图神经网络的节点特征作为输入数据与相应的原位数据进行运算；在混合原位计算模式，对所述图神经网络的所述层将图数据的邻接矩阵和节点特征作为原位数据分别映射到相应的阻变存储器阵列中，以将权重参数作为输入数据与相应的原位数据进行运算。



CN 114186598 A

1.一种用于基于阻变存储器的图神经网络计算装置的图神经网络计算方法,其特征在于,包括:

对于图神经网络的任一层,分析该层中将要在阻变存储器阵列中运算的图数据在权重原位计算模式和混合原位计算模式下的处理时延相对大小,选择时延最小的模式作为该层的计算模式;

在权重原位计算模式,对所述图神经网络的所述层将图数据的邻接矩阵和图神经网络的权重参数作为原位数据分别映射到相应的阻变存储器阵列中,以将图神经网络的节点特征作为输入数据与相应的原位数据进行运算;

在混合原位计算模式,对所述图神经网络的所述层将图数据的邻接矩阵和节点特征作为原位数据分别映射到相应的阻变存储器阵列中,以将权重参数作为输入数据与相应的原位数据进行运算。

2.根据权利要求1所述的方法,其特征在于,对于图神经网络的第一层,其处理时延仅考虑计算时延。

3.根据权利要求2所述的方法,其特征在于,图神经网络的第一层的计算模式通过以下公式确定:

$$t=a^k \times T_{read}\left(\frac{Bitwidth(X^k)}{DAC\_res}\right) - c^k \times T_{read}\left(\frac{Bitwidth(W^k)}{DAC\_res}\right);$$

其中,a^k表示图神经网络的第k层中图数据的节点数或者平均节点数,

$T_{read}\left(\frac{Bitwidth(X^k)}{DAC_res}\right)$表示阻变存储器阵列以第k层节点特征$X^k$作为输入数据时读取第k层的权重值所需的时延,DAC_res表示数模转换的分辨率,c^k表示第k层的权重参数$W^k_{b\times c}$对应维度矩阵的列数c,$T_{read}\left(\frac{Bitwidth(W^k)}{DAC_res}\right)$表示阻变存储器阵列以第k层权重参数$W^k$作为输入信号时读取第k层的节点特征所需的时延,k=1;

其中,t的值大于零则表示该层在混合原位计算模式下的处理时延更小,选择混合原位计算模式,否则,选择权重原位计算模式。

4.根据权利要求1所述的方法,其特征在于,对于图神经网络的其他层,其处理时延考虑计算时延和节点特征更新的写入时延。

5.根据权利要求4所述的方法,其特征在于,图神经网络的其他层的计算模式通过以下公式确定:

$$t' = a^k \times T_{read}\left(\frac{Bitwidth(X^k)}{DAC\_res}\right) - c^k \times T_{read}\left(\frac{Bitwidth(W^k)}{DAC\_res}\right) - \beta;$$

其中,a^k表示图神经网络的第k层中图数据的节点数或者平均节点数,

$T_{read}\left(\frac{Bitwidth(X^k)}{DAC_res}\right)$表示阻变存储器阵列以第k层节点特征$X^k$作为输入数据时读取第k层的权重值所需的时延,DAC_res表示数模转换的分辨率,c^k表示第k层的权重参数$W^k_{b\times c}$对应

维度矩阵的列数c，$T_{read}(\frac{Bitwidth(W^k)}{DAC_res})$表示阻变存储器阵列以第k层权重参数$W^k$作为输入信号时读取第k层的节点特征所需的时延，β表示更新节点特征所需的写入时延，k≠1；

其中，t′的值大于零则表示该层数据在混合原位计算模式下的处理时延更小，选择混合原位计算模式，否则，选择权重原位计算模式。

6.根据权利要求1至5任一项所述的方法，其特征在于，将原位数据映射到相应的阻变存储器阵列包括稀疏数据映射方式和稠密数据映射方式；

其中，稀疏数据映射方式是指将相应原位数据对应的矩阵划分为多个子图，删除其中数值为零的空子图，将非空子图按顺序映射到阻变存储器阵列中；

所述稠密数据映射方式是指将相应原位数据对应的矩阵直接按照阻变存储器阵列大小划分并进行映射。

7.根据权利要求6所述的方法，其特征在于，在进行原位数据映射时，作为原位数据的邻接矩阵均采用稀疏数据映射方式映射到阻变存储器阵列上存储。

8.根据权利要求6所述的方法，其特征在于，在进行原位数据映射时，作为原位数据的权重参数均采用稠密数据映射方式映射到阻变存储器阵列上存储。

9.根据权利要求6所述的方法，其特征在于，针对图神经网络的第一层，在进行原位数据映射时，作为原位数据的节点特征的稀疏度大于预定稀疏阈值的情况下，采用稀疏数据映射方式映射到阻变存储器阵列上存储，否则，采用稠密数据映射方式映射到阻变存储器阵列上存储。

10.根据权利要求9所述的方法，其特征在于，针对图神经网络的其他层，在进行原位数据映射时，作为原位数据的节点特征采用稠密数据映射方式映射到阻变存储器阵列上存储。

11.一种基于阻变存储器的图神经网络计算装置，包括用于图数据的存储和计算的阻变存储器阵列，以及处理单元，其特征在于，所述处理单元被配置为执行权利要求1至10任一项所述的方法。

一种基于阻变存储器的图神经网络计算方法和装置

技术领域

[0001] 本发明涉及神经网络的加速处理技术领域，具体来说涉及图神经网络 在基于阻变存储器的计算装置上的处理技术领域，更具体地说，涉及一种 基于阻变存储器的图神经网络计算方法和装置。

背景技术

[0002] 图学习(Graph Learning)，尤其是图神经网络，是一种用于处理大型 属性图的新兴技术，可以用来表示各种相关的随机数据，如文本、图像和 视频。它已经在社交网络分析、电子商务推荐、生物分子表征等大量的应 用中被证明是有效的。真实世界的图数据(如淘宝上的电子商务数据和微 信上的社交网络数据)通常是大而稀疏的，而面向这些图数据处理的图学 习任务大多是数据密集型的，因此在将其部署在通用处理器上时，会受限 于密集的数据移动瓶颈。

[0003] 基于阻变存储器的加速器通过将选定的静态数据映射到阻变存储器 阵列上来实现数据的原位计算，大幅减少了从内存到通用处理器计算单元 的巨大数据移动开销。因此，它是一种非常适合于数据密集型任务的技术，具有广阔的应用前景。目前在针对神经网络的加速器设计中，基于阻变存 储器的加速器可以达到传统CMOS加速器几个数量级的能效比。因此可以 预见，阻变存储器的存内计算技术同样适合于图学习应用的加速，目前国 内外在这方面的研究尚处于早期阶段。

[0004] 在使用基于阻变存储器的加速器进行计算之前，通常需要预先将一部 分静态数据写在阻变存储器计算阵列中，从而实现原位计算。在已有的基 于阻变存储器的神经网络加速器中，如图1中虚线框内的左侧图像所示，通常将神经网络模型即权重值作为静态数据，预先编程在阻变存储器阵列 中，以进行原位计算，此处称这种模式为权重原位计算模式。它避免了权 重值从内存到计算单元的数据移动。相反，如图1中虚线框内的右侧图像所示称为数据原位计算模式，即将模型输入映射到阻变存储器上。权重原 位计算模式是基于阻变存储器的传统神经网络(例如卷积神经网络)加速 器的最佳选择。原因有以下几点：1.卷积神经网络的输入特征比模型参数 小得多，在内存中保持模型参数的静态在数据移动开销方面比参数的收益 要小。2.卷积神经网络的输入特征值将随着计算过程逐层不断更新，对特 征值的更新将它们不断重复写入阻变存储器单元中，带来巨大的写操作能 耗开销，而权重值不发生改变，因此权重模式只需要较小能耗开销的读操 作。3.在典型的卷积神经网络应用，如图像或语音识别中，很少重用输入 特征值，因此很难从原位特征值数据处理中获益。

[0005] 如以下两个公式所示，卷积神经网络与图学习单层操作存在一些区别：

[0006] 图学习：$X^{l+1}=\sigma(AX^{l}W^{l})$

[0007] 卷积神经网络：$X^{l+1}=\sigma(X^{l}W^{l})$

[0008] 其中，l表示层数，σ()为激活函数。与卷积神经网络不同的是，图学 习单层操作包括两个连续的矩阵乘法与三个矩阵操作数，包括图数据邻接 矩阵A，图特征矩阵X和权重值

矩阵W。这两个矩阵乘法需要分别在阻变 存储器上运算，而两个操作的执行顺序和选择三个操作数中的哪两个操作 数映射到阻变存储器(即原位计算模式的选择)将直接影响到图学习计算 开销。此外，图学习中操作数即矩阵的特点也不同于卷积神经网络。首先， 邻接矩阵A在社交网络等大数据应用中可以是超大稀疏的，在骨架识别、MUTAG数据集等图分类应用中也可以是小规模的。其次，图特征矩阵X本 质上代表了图的特征，在不同的图学习任务中，特征矩阵大小变化范围随 着图大小的变化而变化很大。此外，图学习中的权重值矩阵大小取决于输 入特征维数和输出特征维数，有时可以非常紧凑，如千比特级，甚至比输入图大小还小几个数量级。例如，一些图学习模型不超过三层。最后，与 卷积神经网络不同的是，在一些图学习任务如推荐系统中，输入图特征与 图邻接矩阵存在重用特性，因此适合于作为静态数据映射到阻变存储器阵 列中实现重复调用，从而显著减少数据移动开销。因此在图学习任务中， 权重值并不是主要的数据移动开销，而权重原位计算模式并不总适合于基 于阻变存储器的图学习计算。简而言之，原位计算模式的选择，即决定模 型输入的哪一部分应该作为静态数据存储在阻变存储器阵列中，会导致不 同的数据移动和写操作开销。因此，灵活的静态数据(原位数据)选择对 于基于阻变存储器加速器的图学习任务处理的能效至关重要。

发明内容

[0009] 因此，本发明的目的在于克服上述现有技术的缺陷，提供一种基于阻 变存储器的图神经网络计算方法和装置。

[0010] 本发明的目的是通过以下技术方案实现的：

[0011] 根据本发明的第一方面，提供一种用于基于阻变存储器的图神经网络 计算装置的图神经网络计算方法，包括：对于图神经网络的任一层，分析 该层中将要在阻变存储器阵列中运算的图数据在权重原位计算模式和混 合原位计算模式下的处理时延相对大小，选择时延最小的模式作为该层的 计算模式；在权重原位计算模式，对所述图神经网络的所述层将图数据的 邻接矩阵和图神经网络的权重参数作为原位数据分别映射到相应的阻变存储器阵列中，以将图神经网络的节点特征作为输入数据与相应的原位数 据进行运算；在混合原位计算模式，对所述图神经网络的所述层将图数据 的邻接矩阵和节点特征作为原位数据分别映射到相应的阻变存储器阵列 中，以将权重参数作为输入数据与相应的原位数据进行运算。

[0012] 在本发明的一些实施例中，对于图神经网络的第一层，其处理时延仅 考虑计算时延。

[0013] 在本发明的一些实施例中，图神经网络的第一层的计算模式通过以下 公式确定：

[0014] $$t=a^{k}\times T_{read}\left(\frac{Bitwidth(X^{k})}{DAC\_res}\right)-c^{k}\times T_{read}\left(\frac{Bitwidth(W^{k})}{DAC\_res}\right);$$

[0015] 其中，a^{k}表示图神经网络的第k层中图数据的节点数或者平均节点数，$T_{read}\left(\frac{Bitwidth(X^{k})}{DAC_res}\right)$表示阻变存储器阵列以第k层节点特征$X^{k}$作为输入数据 时读取第k层的权重值所需的时延，DAC_res表示数模转换的分辨率，c^{k}表 示第k层的权重参数$W_{b\times c}^{k}$对

应维度矩阵的列数c，$T_{read}(\frac{Bitwidth(W^k)}{DAC_res})$表示阻 变存储器阵列以第k层权重参数$W^k$作为输入信号时读取第k层的节点特征 所需的时延，k＝1；其中，t的值大于零则表示该层在混合原位计算模式下 的处理时延更小，选择混合原位计算模式，否则，选择权重原位计算模式。

[0016] 在本发明的一些实施例中，对于图神经网络的其他层，其处理时延考 虑计算时延和节点特征更新的写入时延。

[0017] 在本发明的一些实施例中，图神经网络的其他层的计算模式通过以下 公式确定：

[0018] $$t' = a^k \times T_{read}(\frac{Bitwidth(X^k)}{DAC\_res}) - c^k \times T_{read}(\frac{Bitwidth(W^k)}{DAC\_res}) - \beta;$$

[0019] 其中，a^k表示图神经网络的第k层中图数据的节点数或者平均节点数，$T_{read}(\frac{Bitwidth(X^k)}{DAC_res})$表示阻变存储器阵列以第k层节点特征$X^k$作为输入数据 时读取第k层的权重值所需的时延，DAC_res表示数模转换的分辨率，c^k表 示第k层的权重参数$W^k_{b\times c}$对应维度矩阵的列数c，$T_{read}(\frac{Bitwidth(W^k)}{DAC_res})$表示阻 变存储器阵列以第k层权重参数$W^k$作为输入信号时读取第k层的节点特征 所需的时延，β表示更新节点特征所需的写入时延，k≠1；其中，t'的值 大于零则表示该层数据在混合原位计算模式下的处理时延更小，选择混合原位计算模式，否则，选择权重原位计算模式。

[0020] 在本发明的一些实施例中，将原位数据映射到相应的阻变存储器阵列 包括稀疏数据映射方式和稠密数据映射方式；其中，稀疏数据映射方式是 指将相应原位数据对应的矩阵划分为多个子图，删除其中数值为零的空子 图，将非空子图按顺序映射到阻变存储器阵列中；所述稠密数据映射方式 是指将相应原位数据对应的矩阵直接按照阻变存储器阵列大小划分并进 行映射。

[0021] 在本发明的一些实施例中，在进行原位数据映射时，作为原位数据的 邻接矩阵均采用稀疏数据映射方式映射到阻变存储器阵列上存储。

[0022] 在本发明的一些实施例中，在进行原位数据映射时，作为原位数据的 权重参数均采用稠密数据映射方式映射到阻变存储器阵列上存储。

[0023] 在本发明的一些实施例中，针对图神经网络的第一层，在进行原位数 据映射时，作为原位数据的节点特征的稀疏度大于预定稀疏阈值的情况下， 采用稀疏数据映射方式映射到阻变存储器阵列上存储，否则，采用稠密数 据映射方式映射到阻变存储器阵列上存储。

[0024] 在本发明的一些实施例中，针对图神经网络的其他层，在进行原位数 据映射时，作为原位数据的节点特征采用稠密数据映射方式映射到阻变存 储器阵列上存储。

[0025] 根据本发明的第二方面，提供一种基于阻变存储器的图神经网络计算 装置，包括用于图数据的存储和计算的阻变存储器阵列，以及处理单元， 其特征在于，所述处理单元被配置为执行第一方面所述的方法。

附图说明

[0026] 以下参照附图对本发明实施例作进一步说明,其中:

[0027] 图1为根据本发明实施例的数据配置过程中权重原位方式和数据原位 方式的示意图;

[0028] 图2为根据本发明实施例的权重原位计算模式和混合原位计算模式的 示意图;

[0029] 图3为根据本发明实施例的稠密映射和稀疏映射的原理示意图;

[0030] 图4为根据本发明实施例的基于阻变存储器的图神经网络计算装置的 示意图;

[0031] 图5为根据本发明实施例的数据的原位计算模式和映射模式的配置示 意图。

具体实施方式

[0032] 为了使本发明的目的,技术方案及优点更加清楚明白,以下结合附图 通过具体实施例对本发明进一步详细说明。应当理解,此处所描述的具体 实施例仅用以解释本发明,并不用于限定本发明。

[0033] 如在背景技术部分提到的,原位计算模式的选择,会导致不同的数据 移动和写操作开销,灵活的静态数据(原位数据)选择对于基于阻变存储 器加速器的图学习任务处理的能效至关重要。为了实现高能效的阻变存储 器图学习计算,以传统基于阻变存储器的图神经网络计算装置(基于阻变 存储器的图卷积神经网络加速器)中权重原位计算模式为基础,本发明提 出了一种混合原位计算模式,即在阻变存储器上同时采用权重原位方式和数据原位方式进行配置,从而实现图学习一层操作中的两个关键矩阵乘法 计算。此外,不同图学习目标任务具有不同的计算特点,因此适合于不同 的原位计算模式。为了寻找和实践针对不同图学习任务的优化的原位计算 模式,本发明提出了对于图神经网络的每一层,逐层分析将要在阻变存储 器阵列中运算的图数据在权重原位计算模式和混合原位计算模式下的处 理时延相对大小,选择时延最小的模式作为该层的计算模式,相当于一种 任务自适应图学习计算选择策略,用于选择处理时延最小的数据配置方案, 以高效地处理各种图学习计算任务。

[0034] 在对本发明的实施例进行具体介绍之前,先对其中使用到的部分术语 作如下解释:

[0035] 图数据,是指使用节点和边来表示多个节点之间相互关系的图结构。图数据通常存储在图数据库(Graph Database)中。

[0036] 邻接矩阵,用于存储节点间关系的二维数组。

[0037] 权重参数,是指图神经网络模型的模型参数。

[0038] 节点特征,是指节点的特征向量。

[0039] 在图神经网络计算过程中,有两种权重数据,一种是邻接矩阵,另一 种是图神经网络的权重参数(图神经网络模型的参数),以及一种非权重数 据,即节点特征。首先介绍两种原位配置方式,参见图1,DAC(Digital to Analog Converter),表示数字模拟转换器,可将数字信号转换为模拟 电压信号;S&H(Sample and Hold),表示采样保持电路,用于对连续的模 拟信号进行采样,并保存;当对模拟信号进行数字模拟转换时,需要一定 的转换时间,在这个转换时间内,模拟信号要保持基本不变,这样才能保 证转换精度。采样保持电路即为实现这种功能的电路;ADC(Analog to Digital Converter),表示模拟数字转换器,

用于将一个输入模拟信号转 换为一个输出的数字信号;权重原位方式下,是将权重配置为原位数据，将数据(节点特征)作为输入;数据原位方式下,是将数据(节点特征) 配置为原位数据,将权重(邻接矩阵或者权重参数)作为输入。以图1权 重原位方式为例,配置好后,计算过程包括:将代表输入数据(即图神经 网络中的输入激活值)的数字信号输入到DAC里进行数字模拟转换,得到 对应的模拟电压值加到存储计算阵列的每一行上;输入的模拟电压值与阵 列上的阻变存储器根据基尔霍夫原理在列上得到点积和的电流值,该电流 值的大小即输入数据与阵列上预先存储的权重值的矩阵向量积;通过采样 保持电路,对得到的电流结果进行采样并保持不变,方便后续的模拟数字 转换;通过ADC对模拟信号的电流值进行模拟数字转换,得到最终的数字 计算结果。

[0040] 根据以上两种原位配置方式,本发明提出两种不同的原位计算模式，其中,权重原位计算模式是指图神经网络的相应层的计算过程中将两种权 重数据作为原位数据(静态数据)存储在相应的阻变存储器阵列中;混合 原位计算模式是指图神经网络的相应层的计算过程中将一种权重数据和 节点特征作为原位数据存储在相应的阻变存储器阵列中。一个图卷积层中 所有可能的数据流如图2所示,其中,图2a和图2b为两种不同的权重原 位计算模式,图2c和图2d为两种不同的混合原位计算模式。其中,灰色 阵列表示将权重数据作为原位数据存储的阻变存储器阵列,黑色阵列表示 将节点特征作为原位数据存储的阻变存储器阵列,A表示邻接矩阵,W表 示图神经网络的权重参数,X'表示输出的节点特征值,T表示转置。权重 原位计算模式以图2b为例,A和W作为静态数据被预先编程到阻变存储器阵列ReRAM上。当代表输入数据的节点特征X的电压加到存储W的阵列行 上,阵列的列产生X和W矩阵乘法结果XW。然后将XW转置为$(XW)^T$,作 为输入加到存储A的阵列上,得到该层的图卷积结果X'^T。整个过程为$X'=((AXW)^T)^T=((XW)^T(A)^T)^T$。混合原位计算模式下的过程如图2c所示,邻 接矩阵A(代表权重原位)和节点特征X(代表数据原位)被选择为静态数据，映射到基于阻变存储器的数据存储阵列上;其中,首先将转置后的权重W^T输入到存储X^T的阵列中得到$(XW)^T$,然后将其作为新的输入,加到存储A^T的阵列行上进行第二次矩阵乘法,输出得到该图卷积层计算结果为X'^T。整个过程为$X'^T=((AXW)^T)^T=((W^TX^T)(A)^T)$,将$X'^T$再进行转置即可得 到所需的输出的节点特征X'。

[0041] 这里介绍具体的任务自适应的图学习计算模式的选择策略。在提出该 策略之前,申请人对图学习任务进行了分析。因为一个矩阵乘法的两个操 作数不能同时是静态数据,因此,图学习的矩阵计算顺序限制了静态数据 的选择。本发明优选以$A\times(X\times W)$的计算顺序实现图卷积运算,因为 $(A\times X)\times W$的计算顺序将会导致明显的运算次数增加及更多的数据移动开 销。在$A\times(X\times W)$的计算顺序中,只允许同时选择邻接矩阵A和输入特征 值X的混合原位计算模式和邻接矩阵A与权重W的权重原位计算模式。在 这两种模式下邻接矩阵A均默认为静态,所以以下策略的重点在于$(X\times W)$ 的操作中,选择将X还是W作为静态数据映射到数据存储阵列中。由于图 学习模型(图神经网络)中不同的层可能会有不同的特征,因此需要逐层 进行配置,根据本发明的一个实施例,一种图神经网络计算方法,应用于 基于阻变存储器的图神经网络计算装置中,包括:对于图神经网络的每一 层,逐层分析将要在阻变存储器阵列中运算的图数据在权重原位计算模式 和混合原位计算模式下的处理时延相对大小,选择时延最小的模式作为该 层的计算模式;对于选择权重原位计算模式的层,将图数据的邻接矩阵和 图神经网络的权重参数作为原位数据分别映射到不同的阻变存储器

阵列 中，将图神经网络的节点特征作为输入数据与相应的原位数据进行运算；对于选择混合原位计算模式的层，将图数据的邻接矩阵和节点特征作为原 位数据分别映射到不同的阻变存储器阵列中，将权重参数作为输入数据与 相应的原位数据进行运算。优选的，在进行权重原位计算模式时，输入数 据被配置为先与权重参数进行矩阵乘法运算，将得到的结果再与邻接矩阵 进行矩阵乘法运算。该实施例的技术方案至少能够实现以下有益技术效果：本发明同时支持层级的权重原位计算模式与混合原位计算模式，并根据每 层的处理时延支持原位计算模式的自适应选择配置，以高效地处理各种图 学习计算任务。

[0042] 在基于阻变存储器阵列的任务自适应计算模式选择过程中，逐层选择 分为两种类型：第一层选择和其他层选择，因为第一层的全部静态数据均 预先映射到阻变存储器存储阵列中，而其他层在混合原位计算模式下的特 征更新会带来额外的阻变存储器写入开销。

[0043] 由于第一层的静态数据将在计算前预编程到阻变存储器阵列上，而不 需要从主存储器中移动数据，因此不考虑配置的延迟。在原位计算的操作 中，根据阻变存储器原位计算的延迟T自适应选择相应的计算模式。具体 来说，矩阵大小为m×p的A矩阵与大小为p×n的B矩阵进行矩阵乘法运 算时，A原位计算模式的延迟T_A和B原位计算模式的延迟T_B分别定义如下：

[0044] $$T_A = n \times T_{read}\left(\frac{Bitwidth(B)}{DAC\_res}\right)$$

[0045] $$T_B = m \times T_{read}\left(\frac{Bitwidth(A)}{DAC\_res}\right)$$

[0046] 其中，$T_{read}(n)$为阻变存储器读输入为n比特时所需的延迟，Bitwidth(m) 为m的比特位宽，DAC_res为数模转换的分辨率。因此，对于(X×W)这个 操作，比较权重值W为静态数据下的原位计算延迟$T_W = a^k \times T_{read}\left(\frac{Bitwidth(X^k)}{DAC_res}\right)$与输入特征值为静态数据下的原位计算延迟$T_X = c^k \times T_{read}\left(\frac{Bitwidth(W^k)}{DAC_res}\right)$。以下以数模转换的分辨率为1bit为例，则可以转换为 计算$a^k \times T_{read}(Bitwidth(X^k)) - c^k \times T_{read}(Bitwidth(W^k))$。若差值为正，则 说明$T_W > T_X$，此时选择X为静态数据的延迟更小，应采用混合原位计算 模式；若差值为负，则说明$T_W < T_X$，此时选择W为静态数据的延迟更小， 应采用权重原位计算模式。

[0047] 根据本发明的一个实施例，图神经网络的第一层对应的处理时延仅考 虑计算时延。假设数模转换的分辨率为1bit，图神经网络的第一层的计算 模式通过以下公式确定：

[0048] $t = a^k \times T_{read}(Bitwidth(X^k)) - c^k \times T_{read}(Bitwidth(W^k))$；

[0049] 其中，t的值大于零则表示该层在混合原位计算模式下的处理时延更 小，选择混合原位计算模式，否则，选择权重原位计算模式，在本次计算 只有一个图数据时，a^k表示图神经网络的第k层中图数据的节点数或者平 均节点数，$T_{read}(Bitwidth(X^k))$表示阻变存储器阵列以第k层节点特征X^k作 为输入数据时读取第k层的权重值所需的时延，c^k表示第k层的权重参数 $W^k_{b\times c}$对应维度矩阵的列数c，$T_{read}(Bitwidth(W^k))$表示阻变存储器阵列以 第k层权重参数W^k作为输入信号时读取第k层的节点特征所需的时延，k＝1。例如，确定图输入

节点特征为$X^{n}_{a_i\times b}$，i＝1，...，m，图神经网络的权重参 数为$W^{n}_{b\times c}$。其中，n为图神经网络的层数，m为输入图的数量，a_i、b为节 点特征的维度，分别为节点特征对应矩阵的行数a_i和列数b。$W^{n}_{b\times c}$中，b、c 分别为权重参数的维度，分别为权重参数对应矩阵的行数b和列数c。在准 备阶段计算多图任务中多个图输入的平均节点数$a^{k}=\frac{a_1+a_2+\ldots+a_m}{m}$，对于只 有一个大规模图输入的节点型任务(即m＝1时)可以跳过这一阶段。即：在本次计算仅有一个图数据时，a^{k}表示图神经网络的第k层的节点特征 $X^{k}_{a\times b}$对应矩阵的行数a(相当于该图数据的节点数)；在本次计算有多个图 数据时，a^{k}表示图神经网络的第k层多个图数据的节点特征的矩阵的平均 行数(相当于多个图数据的平均节点数)。该实施例的技术方案至少能够实 现以下有益技术效果：对于图神经网络的每一层，处理时延仅考虑计算时 延有助于减少计算模式选择过程的计算量，提高处理效率，降低处理总能 耗。

[0050] 对于其他层，引入一个参数β来衡量由于节点特征更新带来的写入延 迟开销：

[0051] $\beta=\#row\times T_{write}$；

[0052] 其中，#row表示单个阻变存储器阵列总行数，T_{write}表示阻变存储器单 次写的时间。对于其他层，此时输入特征值为静态数据下的原位计算延迟 为 $T_X=c^{k}\times T_{read}(\frac{Bitwidth(W^{k})}{DAC\_res})+\beta$。因此计算$a^{k}\times T_{read}(\frac{Bitwidth(X^{k})}{DAC_res})-c^{k}\times T_{read}(\frac{Bitwidth(W^{k})}{DAC_res})-\beta$的值。若该值为正，则说明$T_W>T_X$，此时选择X为 静态数据的延迟更小，应采用混合原位计算模式；若该值为负，则说明$T_W<T_X$，此时选择W为静态数据的延迟更小，应采用权重原位计算模式。

[0053] 根据本发明的一个实施例，图神经网络的其他层对应的处理时延考虑 计算时延和节点特征更新的写入时延。假设数模转换的分辨率为1bit，图 神经网络的其他层的计算模式通过以下公式确定：

[0054] $t'=a^{k}\times T_{read}(Bitwidth(X^{k}))-c^{k}\times T_{read}(Bitwidth(W^{k}))-\beta$；

[0055] 其中，t'的值大于零则表示该层数据在混合原位计算模式下的处理时 延更小，选择混合原位计算模式，否则，选择权重原位计算模式，k≠1，β表示更新节点特征所需的写入时延。假设图神经网络为3层网络，则此 处k＝2、3。该实施例的技术方案至少能够实现以下有益技术效果：图神经 网络的其他层对应的处理时延考虑计算时延和节点特征更新的写入时延，有助于根据阻变存储器的写入特性考虑相应的时延，以便综合选择最优的 计模式，提高图数据的处理效率。

[0056] 由于图数据可能是稀疏的，如果不加甄别地进行映射，不仅可能导致 配置时延过大，还可能浪费宝贵的阻变存储器资源。根据本发明的一个实 施例，在进行相应原位数据映射时，为该原位数据选择稀疏数据映射方式 或者稠密数据映射方式；其中，稀疏数据映射方式是指将相应原位数据对 应的矩阵划分为多个子图，删除其中数值为零的空子图，将非空子图按顺 序映射到阻变存储器阵列中；所述稠密数据映射方式是指将相应原位数据对应的矩阵直接按照阻变存储器阵列大小划分并进行映射。

[0057] 由于邻接矩阵通常是稀疏的，对邻接矩阵的映射模式进行判别会降低 效率。根据

本发明的一个实施例,在进行原位数据映射时,作为原位数据 的邻接矩阵均采用稀疏数据映射方式映射到阻变存储器阵列上存储。

[0058] 由于图神经网络的权重参数通常是稠密的,对权重参数的映射模式进 行判别会降低效率。根据本发明的一个实施例,在进行原位数据映射时, 作为原位数据的权重参数均采用稠密数据映射方式映射到阻变存储器阵 列上存储。

[0059] 不同于邻接矩阵和权重参数,图神经网络的第一层的节点特征可能是 稀疏的,也可能是稠密的,如果不加甄别的映射,会极大地影响计算效率 和浪费计算资源。根据本发明的一个实施例,针对图神经网络的第一层, 在进行原位数据映射时,作为原位数据的节点特征的稀疏度大于预定稀疏 阈值的情况下,采用稀疏数据映射方式映射到阻变存储器阵列上存储,否 则,采用稠密数据映射方式映射到阻变存储器阵列上存储。优选的,预定稀疏阈值的取值范围是80%～95%。例如,将预定稀疏阈值设为90%。而图神 经网络的其他层的节点特征通常是稠密,根据本发明的一个实施例,针对 图神经网络的其他层,在进行原位数据映射时,作为原位数据的节点特征 采用稠密数据映射方式映射到阻变存储器阵列上存储。

[0060] 举例来说,假设一个矩阵的大小为256*256,数据存储计算阵列大小 为128*128;

[0061] 其中,稠密数据映射方式下:①将256*256的矩阵按照128*128的大 小划分成4块;②分别映射到4块数据存储计算阵列上去。

[0062] 稀疏数据映射方式下:①将256*256的矩阵按照子图大小划分(子图 大小通常为4*4/8*8/16*16,此处以4*4为例),得到4096块;②将4096 小块中块中元素均为零值的块去除,得到剩余的非零块;③将剩余的块按 顺序映射到数据存储计算阵列上去。

[0063] 为了形象地说明,参见图3,以24×24的矩阵和数据存储计算阵列来 说简化说明,长方形表示4×4的矩阵子图,圆形表示4×4个阻变存储器 单元。其中,稠密映射(对应于稠密数据映射方式)时,将24×24的矩阵 直接映射到数据存储计算阵列上去;稀疏映射(对应于稀疏数据映射模式) 时:将24×24的矩阵中元素均为零值的块去除,得到剩余的非零块;将剩 余的块按顺序映射(在阵列上连续映射,中间不留零值的阻变存储器单元) 到矩阵上去。

[0064] 根据本发明的一个实施例,提供一种基于阻变存储器的图神经网络计 算装置,包括用于图数据的存储和计算的阻变存储器阵列,该图神经网络 计算装置被配置为执行前述实施例所述的方法。根据本发明的一个实施例, 参见图4,图神经网络计算装置,包括:数据存储计算模块,包括多个阻 变存储器阵列,所述阻变存储器阵列用于存储原位数据并完成其与输入数 据的矩阵向量乘法。

[0065] 优选的,数据存储计算模块被配置为:

[0066] 支持至少两种计算模式,包括:

[0067] 权重原位计算模式,其中将作为输入数据的节点特征与作为原 位数据的权重参数和邻接矩阵进行矩阵向量乘法;

[0068] 混合原位计算模式,其中将作为输入数据的权重参数与作为原 位数据的节点特征和邻接矩阵进行矩阵向量乘法;

[0069] 对于图神经网络的每一层,根据该层在相应计算模式下的处理时延选 择时延最小的模式作为该层的计算模式。数据存储计算模块包括多个数据 存储阵列构成。

[0070] 根据本发明的一个实施例，图神经网络计算装置还包括：索引控制模 块、多路选择模块、多路选择模块、移位累加模块、激励函数模块、数字 -模拟转换模块、模拟-数字转换模块、寄存器模块、静态随机存储器及其 组合。前面提及的数模转换的分辨率即是指数字-模拟转换模块、模拟-数 字转换模块的分辨率。

[0071] 其中，索引控制模块，用于存储原位数据对应阻变存储器阵列的行、 列地址选通信号，并在计算过程中将对应选通信号输入给数据存储阵列对 应的多路选择模块。

[0072] 多路选择模块(对应于图4中的行多路选择和列多路选择)，根据输 入的选通信号的行、列地址，选通对应的阻变存储器阵列的行与列单元；

[0073] 移位累加模块，用于将输入数据与原位数据进行矩阵向量乘法所得到 的结果移位累加，得到点积计算结果。

[0074] 激励函数模块，用于将所选取的点积计算结果进行激活操作，执行相 应的神经元响应值的计算。

[0075] 数字-模拟转换模块(未在图4示出，可参考图1的DAC)，用于将代 表输入数据(输入激活值)的数字信号转换为对应的模拟电压值，模拟电 压值将加载在对应的阻变存储器阵列行上。

[0076] 模拟-数字转换模块(未在图4示出，可参考图1的ADC)，用于将矩 阵向量乘法所获得的点积计算结果对应的模拟电流值转换为数字值；此处 的点积计算结果可能是经过激励函数模块或者直接由移位累加模块处得 到点积计算结果，具体要看相应的图神经网络对应的任务中是否需要激励 函数模块对相应的点积计算结果进行激活操作；最后一层计算完成后，模 拟-数字转换模块得到的结果为最终的结果数据，表示图数据中各个节点 的嵌入表示(特征向量)。

[0077] 寄存器模块，用于暂时存放即将输入数据存储计算模块的输入数据或 者从数据存储计算模块输出的结果数据。例如，寄存器模块，用于暂时存 放少量即将输入至数据存储计算装置参与计算的输入数据，或已经完成矩 阵乘法操作的输出结果数据。

[0078] 静态随机存储器，用于存放图数据、图神经网络的权重参数和输出的 结果数据。例如，静态随机存储器，用于存放大量即将输入至数据存储计 算装置参与计算的输入数据，或已经完成矩阵乘法操作的输出结果数据。

[0079] 根据本发明的一个实施例，图神经网络计算装置还可以包括处理单元 (图4未示出)，处理单元用于为图神经网络的相应层选择适配的计算模 式，并根据该计算模式配置数据存储计算模块和/或控制配置好的存储计 算模块执行相应图神经网络的运算。处理单元用于控制存储计算模块执行 相应图神经网络的运算、索引控制模块、多路选择模块、多路选择模块、 移位累加模块、激励函数模块、数字-模拟转换模块、模拟-数字转换模块、 寄存器模块、静态随机存储器协同工作以完成相应图神经网络的运算。

[0080] 由以上实施方式可以看出，参见图5，本发明实现了根据图神经网络 和图数据对应的任务自适应选择策略的过程，为图神经网络的每一层(1、 2、……、N层)配置相应的原位计算模式(混合原位或者权重原位)，以 及为所选原位数据根据相应的映射模式进行稀疏映射或者稠密映射；从而 提高图数据的处理效率。基于阻变存储器的图数据的处理过程中，已选定 的原位数据静态映射到阻变存储器阵列中。当输入数据将对应电压信号加 在阵列行上时，即可完成一次图神经网络中某层中的矩阵乘法。阵列的输 出信号将经过模

拟-数字转换模块、移位累加模块和激励函数模块，最终得 到该层的输出，并暂存在静态随机存储器中，用作下一层计算的节点特征。由以上方案可知，本发明的优点在于：本发明在计算模式选择上考虑了图 学习任务的不同计算特点的两个角度：原位计算模式与数据映射策略，结 合具体任务特征有效降低了数据移动能耗开销与存储开销，同时降低了计 算延迟，且对图学习任务计算结果不会有任何影响，实现了高能效的图学 习任务计算。本发明在硬件上支持对原位计算模式与映射方式的灵活配置，从而支持任务自适应的选择策略，为高能效的图学习任务计算提供了硬件 实现。

[0081] 需要说明的是，虽然上文按照特定顺序描述了各个步骤，但是并不意 味着必须按照上述特定顺序来执行各个步骤，实际上，这些步骤中的一些 可以并发执行，甚至改变顺序，只要能够实现所需要的功能即可。

[0082] 本发明可以是系统、方法和/或计算机程序产品。计算机程序产品可以 包括计算机可读存储介质，其上载有用于使处理器实现本发明的各个方面 的计算机可读程序指令。

[0083] 计算机可读存储介质可以是保持和存储由指令执行设备使用的指令 的有形设备。计算机可读存储介质例如可以包括但不限于电存储设备、磁 存储设备、光存储设备、电磁存储设备、半导体存储设备或者上述的任意 合适的组合。计算机可读存储介质的更具体的例子(非穷举的列表)包括：便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可 擦式可编程只读存储器(EPROM或闪存)、静态随机存取存储器(SRAM)、便携式压缩盘只读存储器(CD-ROM)、数字多功能盘(DVD)、记忆棒、软盘、机械编码设备、例如其上存储有指令的打孔卡或凹槽内凸起结构、以及上 述的任意合适的组合。

[0084] 以上已经描述了本发明的各实施例，上述说明是示例性的，并非穷尽 性的，并且也不限于所披露的各实施例。在不偏离所说明的各实施例的范 围和精神的情况下，对于本技术领域的普通技术人员来说许多修改和变更 都是显而易见的。本文中所用术语的选择，旨在最好地解释各实施例的原 理、实际应用或对市场中的技术改进，或者使本技术领域的其它普通技术 人员能理解本文披露的各实施例。

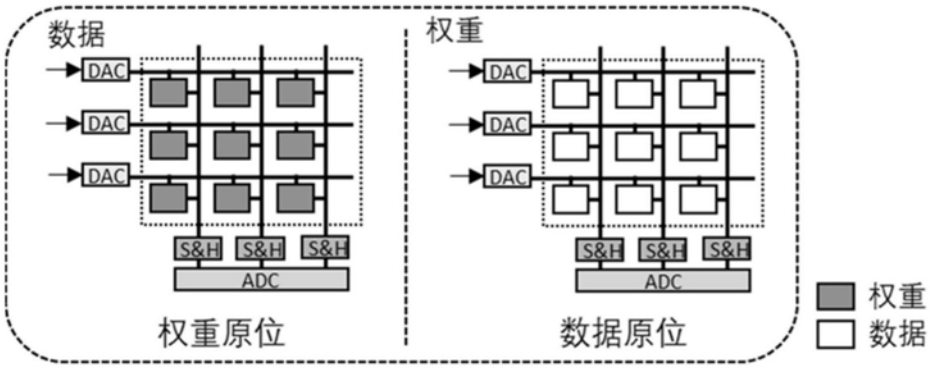

数据
权重
DAC
DAC
DAC
S&H
S&H
S&H
ADC
权重原位
数据原位
权重
数据

图1

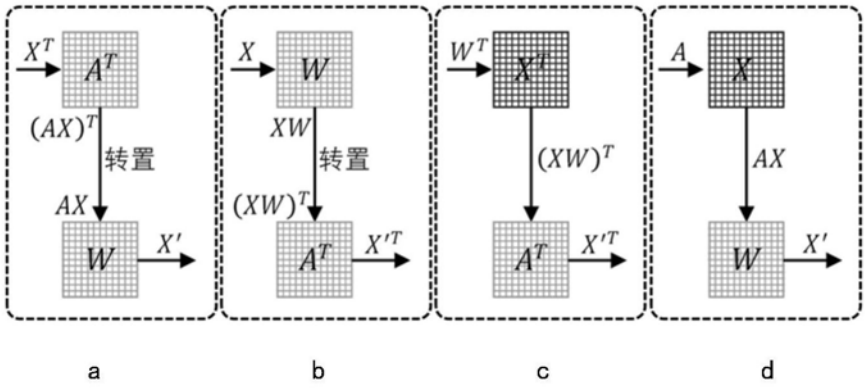

X^T
A^T
$(AX)^T$
转置
AX
W
X'
X
W
XW
转置
$(XW)^T$
A^T
X'^T
W^T
X^T
$(XW)^T$
A^T
X'^T
A
X
AX
W
X'
a
b
c
d

图2

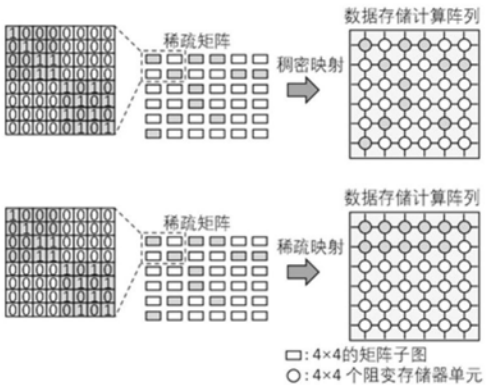

稀疏矩阵
稠密映射
数据存储计算阵列
稀疏矩阵
稀疏映射
数据存储计算阵列
□:4×4的矩阵子图
○:4×4 个阻变存储器单元

图3

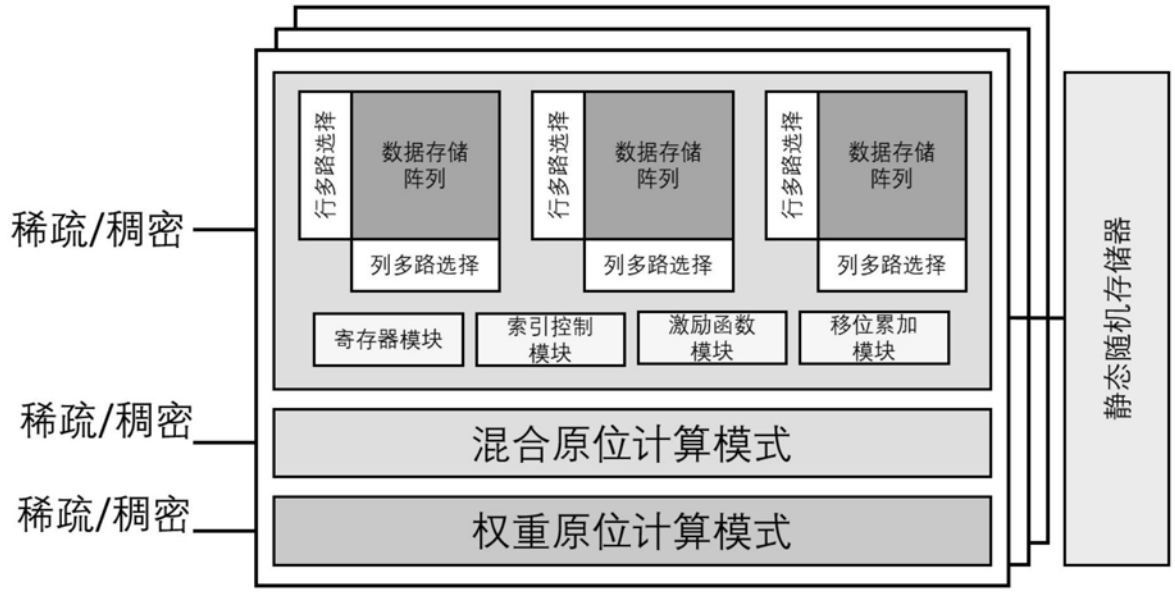

稀疏/稠密
稀疏/稠密
稀疏/稠密
行多路选择
数据存储阵列
列多路选择
行多路选择
数据存储阵列
列多路选择
行多路选择
数据存储阵列
列多路选择
寄存器模块
索引控制模块
激励函数模块
移位累加模块
混合原位计算模式
权重原位计算模式
静态随机存储器

图4

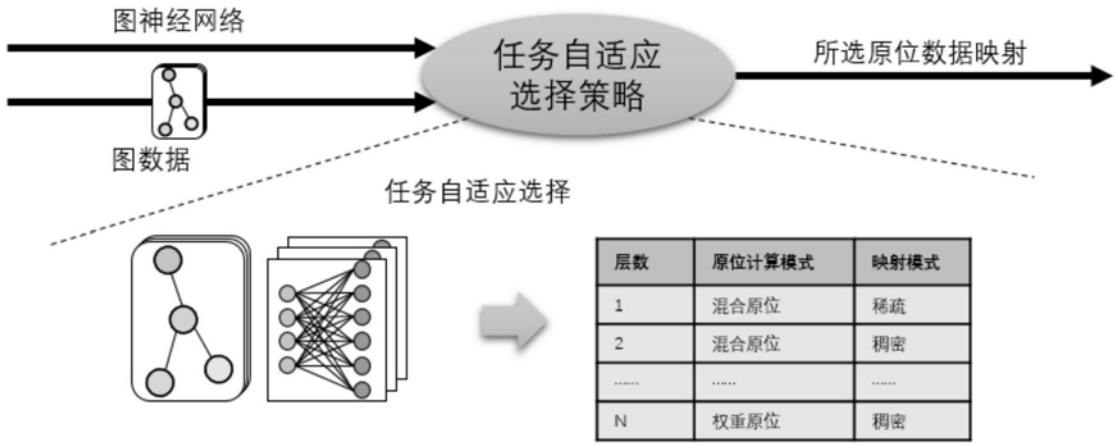

图神经网络
图数据
任务自适应选择策略
所选原位数据映射
任务自适应选择
层数
原位计算模式
映射模式
1
混合原位
稀疏
2
混合原位
稠密
N
权重原位
稠密

图5