**专利申请明细表**

**兹全权委托北京博思佳知识产权代理有限公司办理下列申请专利有关事项**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 发 明 名 称 | | 一种消除冗余的软硬协同的图神经网络加速器 | | |
| 第一  申请人 | 姓名或名 称 | 之江实验室 | | |
| 地址、邮编 | 默认：浙江省杭州市余杭区之江实验室南湖总部，311121  日本专利：之江实验室 中国310023浙江省杭州市余杭区文一西路1818号 | | |
| 企业机构代码或身份证号 | 12330000MB1478604D | | |
| 第二  申请人  （如果为多个申请单位共同申请，可填） | 姓名或名 称 | 华中科技大学 | | |
| 地址、邮编 | 地址：湖北省武汉市洪山区珞喻路1037号，430074 | | |
| 企业机构代码或身份证号 | 12100000441626842D | | |
| 发明人  或设计人 | 姓 名 | 张宇，黄志颖，赵进，余辉，张湛 | 第一发明人身份证号 | 431122198707170515 |
| 专利负责人 | 姓名 | 吴梅英，18662590857，wumeiying@zhejianglab.com  地址：浙江省杭州市余杭区之江实验室南湖总部主楼1402办公室，311100 | | |
| 地址、邮编 |
| 电话/E-mail |
| 技术（撰写）  联系人 | 姓名 | 黄志颖，15180120841，hzying@hust.edu.cn  地址：浙江省杭州市余杭区之江实验室南湖总部11号楼B座B102室，311121 | | |
| 地址、邮编 |
| 电话/E-mail |
| 申请专利类型 | | 🗹发明　　　　　实用新型　　　　外观设计 | | |
| 催缴费联系人电话、Email、地址 | | 吴梅英，18662590857，wumeiying@zhejianglab.com，地址：浙江省杭州市余杭区之江实验室南湖总部主楼1402办公室，311100 | | |

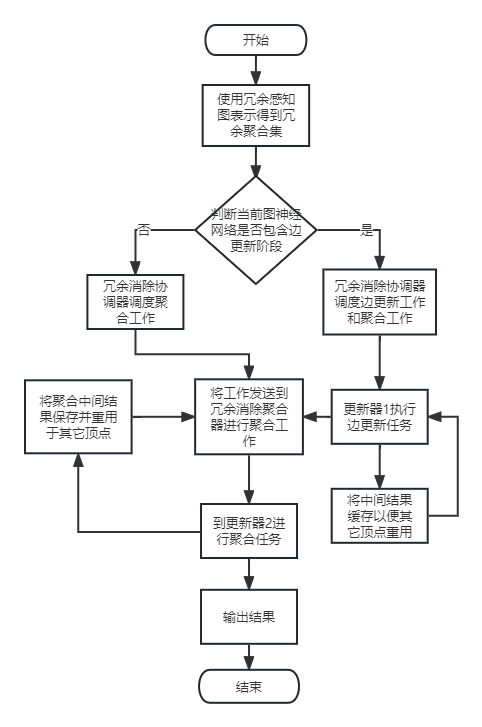
注：

1、红色字体部分需要填写或选择

说明书摘要

本发明实现了一种用于加速图神经网络的软硬协同加速器，该加速器由算法和架构协同设计组成。本发明首先提出了一种用于加速器的动态冗余消除邻域消息传递算法。然后设计了一种新颖的体系结构来支持动态冗余消除邻域消息传递算法并将冗余消除转化为性能改进。同时该加速器是一种可配置的流水线架构，可以配置支持不同的图神经网络变体，并可以保证负载均衡。

摘要附图



权利要求书

1. 一种消除冗余的软硬协同的图神经网络加速器，包括以下部分：

（1）软件部分。一种用于图神经网络的动态消除冗余消息传递算法，通过冗余感知图表示识别冗余聚合集，重用边更新和聚合阶段的中间结果。

（2）硬件部分。设计了一种新型架构的图神经网络加速器，包括冗余消除协调器、冗余消除聚合器、更新器、中央控制器。可以有效支持所提出的动态冗余消除消息传递算法并减少计算不规则性。

1. 如权利要求1所述的方法，其特征在于，所述步骤（1）包括如

下子步骤：

（2.1）动态调度阶段。通过遍历所有冗余集邻居和原始邻居以及基于其状态信息的每个冗余集邻居中的所有原始顶点来调度边更新工作和聚合工作。

在边更新时，如果没有存储遍历顶点的边更新中间结果，则这些顶点将被赋值并保留到顶点列表中，用于后面的边更新。如果已经存储了遍历顶点的边更新中间结果，那么后面的聚合只需要取遍历顶点的中间结果。

在聚合时，对于每个冗余集邻居，对每个原始顶点的所有边更新结果进行聚合。每个目标顶点需要对所有冗余集邻居和原始邻居进行聚合。聚合工作负载保留在工作列表中。

（2.2）冗余消除边更新阶段。根据边权值对顶点列表中的每个顶点进行边更新。由于在调度阶段只有没有进行边更新操作的顶点被调度到顶点列表中，这部分边更新可以被消除。该阶段的结果与工作列表中现有的工作负载相结合，形成新的聚合工作。同时，本阶段将每条边进行边更新的结果缓存起来，以便在对其他目标顶点执行边更新时重复使用。

（2.3）冗余消除聚合阶段。对工作列表中的每项工作进行聚合。工作列表中的每个工作包含一组需要聚合的顶点。该阶段的结果与工作列表中已有的工作相结合，形成新的聚合工作。同时缓存每个聚合集的中间结果。在执行包含这些聚合集的其他目标顶点期间，可以获取和重用结果。因此，消除了冗余聚合集的聚合冗余。

（2.4）顶点更新阶段。结合每个目标顶点的旧顶点特征和聚合特征，以根据顶点权重产生新的顶点特征。

1. 如权利要求2所述的方法，其特征在于，所述步骤（2）包括如

下子部分：冗余消除协调器、冗余消除聚合器、更新器、中央控制器。

（3.1）中央控制器：配置这些引擎的流水线执行流，以支持各种图神经网络。这些引擎的执行流水线符合所提出的动态冗余消除消息传递算法，该算法可以在一个阶段消除冗余。它由一个中央控制器和六个可变配置的控制器组成。

（3.2）冗余消除协调器：冗余消除调度器通过遍历所有冗余集邻居和原始邻居，以及每个冗余集邻居中的所有原始顶点，根据它们的状态信息，将边更新工作分派给更新器，并将聚合工作分派给冗余消除聚合器。包含一个图结构预取器、一个冗余消除调度器和一个顶点预取器，其中图结构预取器从高带宽内存预取结构和状态信息，冗余消除调度器调度不规则计算，而顶点预取器将顶点属性预取到相应的缓存器中以供执行。

（3.3）冗余消除聚合器：支持本发明提出的冗余消除聚合算法，该算法可以执行目标顶点和聚合冗余集的聚合任务。它包含一个邻居缓存器、一个聚合缓存器和一个聚合引擎。邻居缓存器使冗余消除聚合器能够重用原始顶点特征、边更新的中间结果和聚合的冗余聚合集，以减少实时通信冗余。邻居缓存器将不规则的聚合任务组装成规则的聚合任务，并由聚合引擎执行。此外，它可以缓存聚合中间结果，以便这些中间结果可以重新用于其他顶点的后续执行。

（3.4）更新器，本发明中加速器又两个更新器以支持图神经网络变体。对于具有边更新阶段的图神经网络变体，更新器1执行边更新，而更新器2执行顶点更新。更新器1接收并执行冗余消除聚合器派发的更新任务。此外，更新器支持多粒度使用。

（4）缓存器管理：为了减少数据访问的延迟，本发明使用片上缓存器来缓存各种数据和中间结果，以提高数据重用率。除了更新缓存器和聚合缓存器，还有权重缓存器、结构缓存器和即时缓存。

说明书

一种消除冗余的软硬协同的图神经网络加速器

**技术领域**

本发明属于大数据处理中的图学习技术领域，具体涉及一种图神经网络加速器的设计方法，利用消除图神经网络涉及到的三个阶段：边更新，聚合和顶点更新中的计算冗余来加快图神经网络训练，训练精度保持不变

**背景技术**

图神经网络是基于图结构化数据的深度学习技术，在各种基于图的任务中取得了卓越的性能，例如计算机视觉、自然语言处理和推荐系统。现有的典型图神经网络主要基于递归邻域消息传递方法，其中每个顶点接收其邻居的特征，并使用接收到的消息在图神经网络的单层内更新自己的特征。递归实现多邻域聚合，以便顶点可以从其k-hop网络邻域中的其他顶点收集信息。最终特征用于下游预测任务，包括图分类、节点分类和链接预测。

扩展图神经网络以有效处理大型真实世界图数据集仍然是一个巨大的挑战。例如，电子商务图数据包含数十亿用户和项目。特别是，许多实现旨在处理整个图拉普拉斯矩阵，即使对于中等大小的图，它也无法放入主内存。为解决此问题，现有技术使用图神经网络的顶点程序抽象将消息传递层的数据流抽象为三个重要阶段：边更新、聚合和顶点更新，一般来说，边更新通过利用神经网络计算每条边的消息，而聚合为每个目标顶点收集相邻消息。最终，顶点更新结合每个目标顶点的旧顶点特征和聚合特征，使用神经网络生成每个目标顶点的新顶点特征。图神经网络的不同变体具有这些阶段的不同组合。边更新和顶点更新阶段和传统神经网络一样，通常由矩阵向量乘法表示。聚合阶段包含了大部分图处理行为，它严重依赖于本质上随机和稀疏的图结构。每个目标顶点的处理需要聚合来自其所有源邻居的特征。

这些计算过程包括常规计算和大量不规则的执行模式，不利于使用CPU和GPU。专用硬件架构是提高图神经网络计算效率的一个有效的选择。过去的图处理器和图神经网络加速器只能优化以支持图处理或图神经网络，不能同时支持两者。为此，一些遵循邻域消息传递机制的计算抽象的定制加速器被提出，与GPU和现有的图神经网络和深度学习加速器相比，在降低能耗的同时实现了更高的性能。上述加速器利用顶点程序抽象并独立地对每个顶点进行邻域聚合。但是，这种为每个顶点独立进行邻域消息传递的方式会为边更新和聚合引入大量冗余计算。因为当顶点共享相同的邻居时，这些邻居在边更新或聚合阶段中为特定顶点生成的中间结果可以重复用于具有相同邻居的其他顶点。现有的图神经网络加速器忽略了图神经网络的这一特性，存在大量冗余计算。同时每个顶点的特征维数一般都很大，例如：512、1024、2048。因此，边更新计算非常耗时，将顶点特征从片外存储器传输到片上存储器的开销也很大。

**发明内容**

针对现有技术的不足，本发明提出一种消除冗余的图神经网络加速器，该加速器是软硬协同设计支持，包含一种用于图神经网络的动态消除冗余消息传递算法，一种新颖的架构设计和种冗余感知动态调度机制，以在一个阶段消除边更新和聚合的计算冗余。所提出的算法可以识别和消除计算冗余，执行与传统图神经网络相同的计算并提供相同的模型/精度；架构设计帮助识别和消除冗余计算并进一步减少冗余通信；冗余感知动态调度机制在运行时安排冗余消除的任务，有效地解决计算不规则性并通过高度并行化的计算引擎利用任务内规律性。

为实现上述目的，本发明中动态消除冗余消息传递算法包括以下步骤：

（1）动态调度阶段。通过遍历所有冗余集邻居和原始邻居以及基于其状态信息的每个冗余集邻居中的所有原始顶点来调度边更新工作和聚合工作。在边更新时，如果没有存储遍历顶点的边更新中间结果，则这些顶点将被赋值并保留到顶点列表中，用于后面的边更新。如果已经存储了遍历顶点的边更新中间结果，那么后面的聚合只需要取遍历顶点的中间结果。在聚合时，对于每个冗余集邻居，该阶段需要对每个原始顶点的所有边更新结果进行聚合。对每个目标顶点，需要对所有冗余集邻居和原始邻居进行聚合。聚合工作负载保留在工作列表中。

（2）冗余消除边更新阶段。根据边权值对顶点列表中的每个顶点进行边更新。由于在调度阶段只有没有进行边更新操作的顶点被调度到顶点列表中，这部分边更新可以被消除。该阶段的结果与工作列表中现有的工作负载相结合，形成新的聚合工作。同时，将每条边进行边更新的结果缓存起来，以便在对其他目标顶点执行边更新时重复使用。

（3）冗余消除聚合阶段。对工作列表中的每项工作进行聚合。工作列表中的每个工作也是一个列表，其中包含一组需要聚合的顶点。该阶段的结果与工作列表中已有的工作相结合，形成新的聚合工作。同时缓存每个聚合集的中间结果。在执行包含这些聚合集的其他目标顶点期间，可以获取和重用结果。因此，消除了冗余聚合集的聚合冗余。

（4）顶点更新阶段。结合每个目标顶点的旧顶点特征和聚合特征，以根据顶点权重产生新的顶点特征。

本发明中，冗余聚合集表示一组公共的邻居。如果这些冗余聚合集的中间结果可以在第一次聚合后存储，则聚合结果可以在后续的消息传递过程中重复使用。

冗余感知图结构：在图中，一个目标顶点有一些相邻的顶点。将所有冗余聚合集视为独立的邻居，以替换相应冗余聚合集中的所有顶点。因此，V中顶点的邻居可以分为两类：冗余聚合集和原始顶点。一个冗余聚合集包含一组原始顶点。为了区分冗余聚合集和原始顶点，每个冗余聚合集或每个原始顶点都与一个ID号和ID类型相关联。ID类型用bit来表示是冗余聚合集或原始顶点。为了识别边更新和聚合冗余，每个原始顶点都与一个状态位相关联，它表示原始顶点是否已经进行了边更新操作。每个冗余聚合集还关联一个状态位，表示该冗余聚合集是否进行过聚合操作。

冗余感知批量执行是指利用顶点间隔和边分片的思想来划分图结构，批量执行特定分区。因此，这个块中的顶点特征只需要从片外存储器提取一次到片上存储器。这减少了计算所需的片上内存的峰值量，因为只有一部分图需要加载一次。此外，在消息传递过程中，特定顶点的边更新操作和特定冗余聚合集的聚合操作被执行一次，然后被该批量中的所有目标顶点重复使用。

本发明中图神经网络加速器架构为：包括冗余消除协调器、冗余消除聚合器、更新器、中央控制器。

（1）中央控制器：配置这些引擎的流水线执行流，以支持各种图神经网络。这些引擎的执行流水线符合所提出的动态冗余消除消息传递算法，该算法可以在一个阶段消除冗余。它由一个中央控制器和六个可变配置的控制器组成。中央控制器可以异步发出控制命令来管理所有可配置的开关。这些可变配置的控制器可以根据控制命令进行配置，以匹配目标执行流程。

（2）冗余消除协调器：识别计算冗余，将冗余感知边更新和聚合任务的不规则计算解耦并组装到多个常规边更新和聚合任务中，以利用高度并行计算引擎的任务内规律性。具体来说，只有没有进行边更新操作的顶点被发送到更新器执行。在聚合任务中，只有针对未聚合的目标顶点和聚合冗余集的任务才会发送给冗余消除聚合器执行。包含一个图结构预取器、一个冗余消除调度器和一个顶点预取器，其中图结构预取器从高带宽内存预取结构和状态信息，冗余消除调度器调度不规则计算，而顶点预取器将顶点属性预取到相应的缓存器中以供执行。冗余消除调度器通过遍历所有冗余集邻居和原始邻居，以及每个冗余集邻居中的所有原始顶点，根据它们的状态信息，将边更新工作分派给更新器，并将聚合工作分派给冗余消除聚合器。

（3）冗余消除聚合器：支持本发明提出的冗余消除聚合算法，该算法可以执行目标顶点和聚合冗余集的聚合任务。它包含一个邻居缓存器、一个聚合缓存器和一个聚合引擎。邻居缓存器使冗余消除聚合器能够重用原始顶点特征、边更新的中间结果和聚合的冗余聚合集，以减少实时通信冗余。邻居缓存器将不规则的聚合任务组装成规则的聚合任务，存储在聚合任务缓存器中。聚合引擎在聚合任务缓存器中获取聚合任务并相应地执行。此外，它可以配置为将冗余聚合集的聚合中间结果存储到缓存或DRAM中，以便这些中间结果可以重新用于其他顶点的后续执行。

（4）更新器，本发明中加速器又两个更新器以支持图神经网络变体。对于具有边更新阶段的图神经网络变体，更新器1执行边更新，而更新器2执行顶点更新。更新器1接收并执行冗余消除聚合器派发的更新任务。如果它是针对边更新的，可以将其配置为存储中间的边更新结果。此外，更新器支持多粒度使用，它们可以组合在一起用于顶点更新，以支持没有边更新的图神经网络，并进一步避免更新器闲置。本发明在更新器中采用脉动数组进行矩阵乘法操作

（5）缓存器管理：为了减少数据访问的延迟，本发明使用片上缓存器来缓存各种数据和中间结果，以提高数据重用率。除了更新缓存器和聚合缓存器，还有权重缓存器、结构缓存器和即时缓存。权重缓存器用于存储图神经网络的权重，结构缓存器用于存储结构和状态信息，而即时缓存用于存储冗余消除聚合器和更新器生成的中间结果。为了隐藏访问延迟，采用双缓冲技术，允许对所有缓存器重叠执行不同的操作。

总体而言，通过本发明所构思的以上技术方案与现有技术相比，能够取得下列有益效果：

（1）大幅度减少冗余计算和冗余通信，本发明提供的动态冗余消除消息传递和加速器中结构设计可以在聚合阶段重用中间结果，避免了冗余聚合和冗余聚合集中元素的DRAM访问，还可以在边更新阶段避免具有边更新的图神经网络变体执行冗余的边更新计算。并有负载控制器均衡更新器1和更新器2之间的负载。

（2）减少内存访问次数，本发明可以重用用冗余聚合集的中间聚合结果，并将对元素的多个顶点特征的DRAM访问替换为集合的一个单一聚合特征。可以显著减少内存访问次数，内存访问次数随着聚合操作次数的减少而减少。

（3）降低能耗，通过减少边更新工作负载和分别在它们上执行的冗余聚合来节省更新器和冗余消除聚合器的能量。其次，该加速器节省了DRAM的能量，因为本发明所提出的批量执行和缓存管理技术通过利用图局部性提高了DRAM带宽利用率，并且在聚合阶段消除冗余的聚合减少了对顶点特征的片外内存访问次数。

**附图说明**

图1为本发明动态消除冗余消息传递算法流程图。

图2为本发明加速器运行流程图。

图3为本发明硬件加速器结构设计图。

**具体实施方式**

为了使本发明的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本发明进行进一步详细说明。应当理解，此处所描述的具体实施例仅用以解释本方面，并不用于限定本发明。此外，下面所描述的本发明的各个实施方式中所涉及到的技术特征只要彼此之间未构成冲突就可以相互组合。

本发明提供的一种消除冗余的软硬结合的图神经网络加速器，软件部分为动态消除冗余消息传递算法，主要思想是：通过一种图感知表示，识别顶点之间的公共邻居，设为冗余聚合集，通过保存边更新和聚合阶段的中间结果来消除图神经网络中消息传递过程的边更新和聚合冗余，并最小化通信冗余。此外，本发明提出的算法可以减轻边缘更新和聚合计算的不规则性，从而提高进程间和进程内的并行性。硬件部分支持算法的加速器架构。

（1）冗余消除协调器

（1.1）结构预取器将权重、图结构信息提取到缓存器。

（1.2）冗余消除调度器根据结构缓存器中缓存的结构，遍历有冗余集邻居和原始邻居，以及每个冗余集邻居中的所有原始顶点，根据它们的状态信息来调度边更新任务和聚合任务。工作负载被插入到工作先入先出队列中，其中包含顶点属性的确切源地址和目标地址，例如冗余消除聚合器或更新器1中的地址。

（1.3）顶点预取器负责从高带宽缓存中获取顶点属性。

（2）顶点聚合器支持本发明提出的冗余消除聚合算法，并批量重用原始顶点特征、边更新和聚合阶段的中间结果，以减轻即时通信冗余。冗余消除聚合器工作包括两个阶段。

（2.1）一个工作负载中的每个邻居特征（包括冗余邻居集的即时缓存）仅从DRAM中加载一次，然后所有的邻居特征都被邻居缓存器缓存并重新加载到聚合任务缓存器中的多个任务槽中。每个任务槽被分配给一个聚合任务，该任务可以聚合冗余聚合集或目标顶点。

（2.2）邻居缓存区通过重用批次内特征消除了通信冗余。它还将特征缓存并对齐到多个任务槽中，以实现聚合引擎的高度并行执行。每个具有任务ID的聚合任务都分配给一个任务槽。每个任务槽都包含一个存储分配任务ID的目标ID，一个存储相邻特征的特征缓冲区和一个表示该任务是否准备就绪的状态。一个N×M的横杆负责根据任务ID和输出端口之间的映射，将属于聚合任务的邻居特征洗牌到其分配的任务槽。N个输入端口连接构成邻居特征源的冗余消除聚合器、更新器和聚合引擎，而M个输出端口连接M个独立的任务槽。各种任务的输入邻居特征在分配的任务槽中进行缓存和对齐。属于公共邻居的特征将被复制并重新用于相应的插槽。在聚合任务的所有邻居特征都被缓存到它的任务槽之后，这个任务槽被设置为它已经准备好并准备好聚合引擎聚合。

（2.3）在聚合阶段，聚合引擎首先轮询冗余聚合集的任务槽并执行聚合。然后将聚合结果发送到邻居缓存区并缓存到具有此设置邻居的目标顶点的插槽中，以供后续执行。这种机制确保每个冗余聚合集只聚合一次，然后中间结果可以重新用于此工作负载中的目标顶点。同时，中间结果也被发送到缓存和DRAM，这使得它们可以被其他目标顶点重用。一旦所有槽冗余聚合集聚合完毕，聚合引擎将继续聚合特定目标顶点的剩余槽。请注意，尽管特定目标顶点的整个聚合是不规则的，但每个任务槽中的特征是对齐的，这有助于高度并行的聚合引擎利用任务内的规律性。

（2.4）聚合引擎包含一个聚合调度器、一个A-PE组和一个目标比较器。聚合调度器保证管道执行并将工作负载分配以任务分散聚合模式工作，用于工作负载平衡和任务级并行。聚合调度器首先轮询任务槽的状态，选择一批准备好的任务并将特征发送到A-PE组进行聚合。工作负载的数量取决于特征的维度和A-PE组的计算能力。A-PE组完成工作量后，将根据任务类型传输聚合结果。由于目标顶点和冗余聚合集的ID属于不同的数值区间，本发明设计了一个任务比较器来评估任务类型并确定将结果发送到哪里。

（2.5）更新器，无论是边更新还是顶点更新，都利用多层感知器来更新顶点特征采用经典脉动数组构造两个更新器如图3所示，更新器可以配置他们的执行模式用控制器来支持两个主要的图神经网络分类，并做到负载均衡没有空闲资源。

带有边更新的图神经网络的执行模式。如图3所示，更新器独立工作以支持不同的更新操作，每个更新器单独执行矩阵向量乘法操作。在这种情况下，更新器1执行边更新，其中顶点集的原始特征向量连同边更新的权重参数矩阵一起转移到矩阵向量乘法脉动数组中。更新器2执行顶点更新，这也是一组顶点的聚合特征向量与顶点更新的权重参数矩阵之间的矩阵向量乘法操作。在这种模式下，权重参数在更新器之间共享，并且仅在模块内重用。可以配置它来存储更新器1的中间结果，以便这些中间结果可以重新用于其他目标顶点的后续执行。

没有边更新的图神经网络的执行模式。对于没有边更新的图神经网络变体，当更新器1停止运行时，更新器1将处于空闲状态。尽管如此，除了单独工作之外，这些收缩模块还可以进一步组装在一起形成一个大型收缩模块，使更多的元素能够同时执行矩阵向量乘法。利用没有边更新的图神经网络变体的这一特性，利用控制器4来组合更新器1和更新器2，它们组成一个大的更新器来执行顶点更新，如图3所示。这样，所有的更新器都可以重用权重参数。此外，与上面提到的独立工作模型不同，可以将更多的顶点转移到组合的更新器中并行处理。

（2.6）可配置缓存管理，为了进一步减少片外内存通信，本发明设计了一种缓存管理方案，将顶点属性缓存在片上内存中。本发明设计的的缓存管理方案可配置为支持不同类型的图神经网络变体。(i)对于具有边冗余的的图神经网络变体，该方案使用用于存储边更新的中间结果的顶点缓存。对于没有边冗余的变体，该方案缓存原始顶点属性。(ii)对于具有顶点冗余集的图神经网络变体，该方案还使用顶点缓存来存储冗余集的中间聚合结果。为了提高缓存命中率，本发明同时考虑了顶点度信息和冗余集中元素的数量。一些缓存条目是为高优先级顶点保留的。原始顶点的优先级被设置为顶点的度，而冗余集的优先级是其元素的数量。

本专利对于现有图神经网络均表现出良好的性能，以下以图卷积网络为例，详细描述本发明的实施过程：图卷积网络是基于图机器学习的非常强大的神经网络体系结构。本发明提出的一种消除冗余的软硬协同的图神经网络加速器，软件部分可以通过缓存中间结果在一个阶段消除边更新和聚合的计算冗余。所提出的算法可以识别和消除计算冗余，执行与传统图神经网络相同的计算并提供相同的模型/精度。但它会导致严重的不规则计算和内存模式，基于此，本发明设计了硬件，可以有效支持高效冗余消除的图神经网络加速器，可以有效支持所提出的动态冗余消除消息传递算法，并进一步缓解不规则性计算和内存模式。该加速器大大减少了图卷积网络各个阶段的冗余计算和通信。首先，在聚合阶段，加速器通过将其中间聚合结果重用图卷积网络模型，避免了冗余聚合和冗余聚合集中元素的 DRAM 访问。并将对元素的多个顶点特征的 DRAM 访问替换为集合的一个单一聚合特征，减少了内存访问次数；其次，在边更新阶段，本发明加速器避免了源顶点执行冗余的边更新计算。第三，所提出的批量执行方案和可配置的缓存管理进一步提高了数据重用并节省了 DRAM 带宽。最后，重新协调器处理消除冗余消息传递中的不规则工作负载，并支持细粒度的引擎间管道，提升图神经网络计算效率。

本领域的技术人员容易理解，以上所述仅为本发明的较佳实施例而已，并不用以限制本发明，凡在本发明的精神和原则之内所作任何修改，等同替换和改进等，均应包含在本发明的保护范围之内。

**说明书附图**

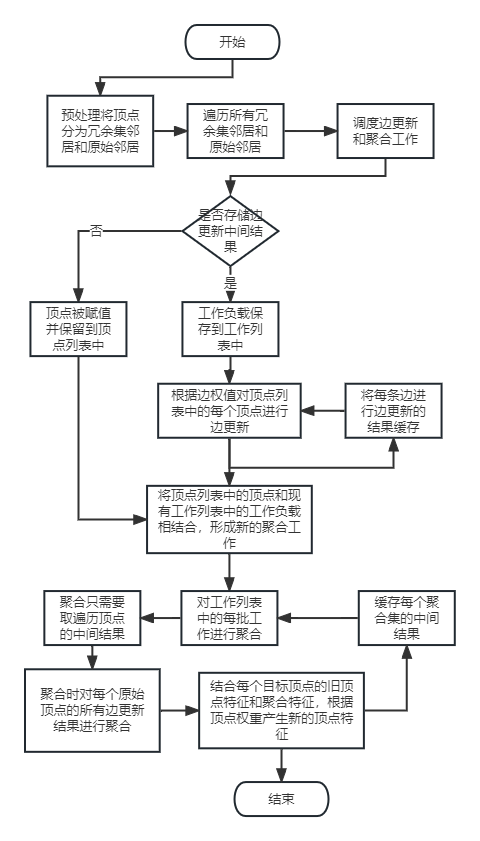


图1

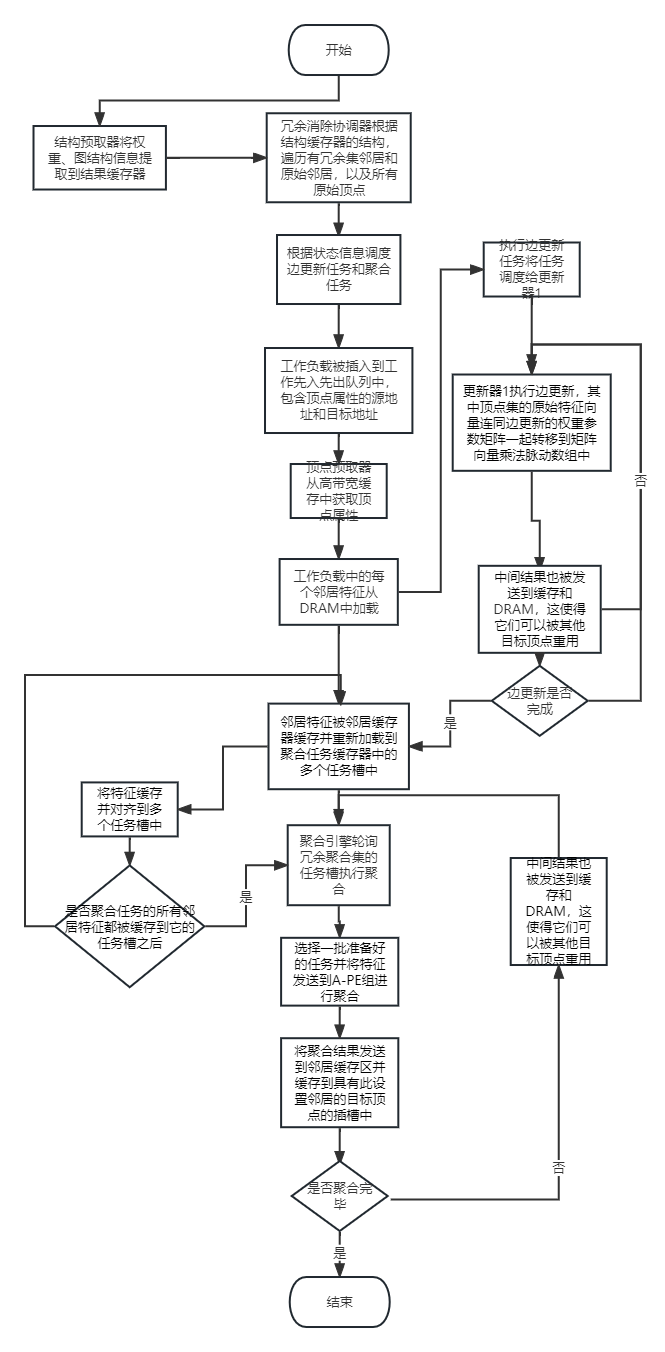


图2



图3