(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 111079924 B
(45)授权公告日 2021.01.08

(21)申请号 201811220957.1
(22)申请日 2018.10.19
(65)同一申请的已公布的文献号
申请公布号 CN 111079924 A
(43)申请公布日 2020.04.28
(73)专利权人 中科寒武纪科技股份有限公司
地址 100190 北京市海淀区科学院南路6号科研综合楼644室
(72)发明人 不公告发明人
(74)专利代理机构 北京林达刘知识产权代理事务所(普通合伙) 11277
代理人 刘新宇
(51)Int.Cl.
G06N 3/08(2006.01)
G06N 3/063(2006.01)

审查员 宋泽宇

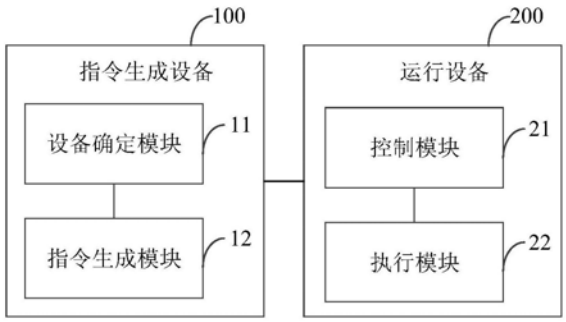
权利要求书6页 说明书23页 附图6页

(54)发明名称

运算方法、系统及相关产品

(57)摘要

本公开涉及一种运算方法、系统及相关产品。该系统包括指令生成设备和运行设备。指令生成设备包括:设备确定模块用于根据接收到的宏指令,确定执行宏指令的运行设备;指令生成模块用于根据宏指令和运行设备,生成运行指令。运行设备包括:控制模块用于获取所需数据、神经网络模型以及运行指令,对运行指令进行解析,获得多个解析指令;执行模块用于根据所述数据执行所述多个解析指令,得到执行结果。本公开实施例所提供的运算方法、系统及相关产品,可跨平台使用,适用性好,指令转换的速度快、处理效率高、出错几率低,且开发的人力、物力成本低。



CN 111079924 B

1.一种神经网络指令处理系统,其特征在于,所述系统包括指令生成设备和运行设备,

所述指令生成设备,包括:

设备确定模块,用于根据接收到的宏指令,确定执行所述宏指令的运行设备;

指令生成模块,用于根据所述宏指令和所述运行设备,生成运行指令;

所述运行设备,包括:

控制模块,用于获取所需数据、神经网络模型以及所述运行指令,对所述运行指令进行解析,获得多个解析指令;

执行模块,用于根据所述数据执行所述多个解析指令,得到执行结果。

2.根据权利要求1所述的系统,其特征在于,所述指令生成设备还包括:

资源获取模块,用于获取备选设备的资源信息,

所述设备确定模块,包括以下至少一个子模块:

第一确定子模块,用于在确定所述宏指令中包含指定设备的标识,且所述指定设备的资源满足执行所述宏指令的执行条件时,将所述指定设备确定为所述运行设备;

第二确定子模块,用于在确定所述宏指令中不包含所述指定设备的标识时,根据接收到的宏指令和所述备选设备的资源信息,从所述备选设备中确定出用于执行所述宏指令的运行设备;

第三确定子模块,在确定所述宏指令中包含所述指定设备的标识,且所述指定设备的资源不满足执行所述宏指令的执行条件时,根据所述宏指令和所述备选设备的资源信息,确定运行设备,

其中,所述执行条件包括:所述指定设备中包含与所述宏指令相对应的指令集,所述资源信息包括所述备选设备所包含的指令集。

3.根据权利要求2所述的系统,其特征在于,所述宏指令包含输入量和输出量中的至少一项,

所述指令生成模块,还用于确定所述宏指令的数据量,根据所述宏指令的数据量、所述宏指令和所述运行设备的资源信息,生成运行指令,

其中,所述数据量是根据所述输入量和所述输出量中的至少一项确定的,所述运行设备的资源信息还包括存储容量、剩余存储容量的至少一项。

4.根据权利要求3所述的系统,其特征在于,所述指令生成模块,包括以下至少一个子模块:

第一指令生成子模块,用于在确定所述运行设备为一个,且所述运行设备的资源不满足执行所述宏指令的容量条件时,根据所述运行设备的运行数据量和所述数据量将所述宏指令拆分成多条运行指令,以使所述运行设备依次执行多条运行指令,

第二指令生成子模块,用于在确定所述运行设备为多个时,根据每个运行设备的运行数据量和所述数据量对所述宏指令进行拆分,生成对应于每个运行设备的运行指令,

其中,运行设备的运行数据量是根据运行设备的资源信息确定的,所述运行指令包含运行输入量和运行输出量中的至少一项,所述运行输入量和所述运行输出量是根据执行所述运行指令的运行设备的运行数据量确定的。

5.根据权利要求1所述的系统,其特征在于,所述指令生成设备还包括:

队列构建模块,用于根据队列排序规则对所述运行指令进行排序,根据排序后的运行

指令构建与所述运行设备相对应的指令队列。

6.根据权利要求2所述的系统,其特征在于,所述指令生成设备还包括:

宏指令生成模块,用于接收待执行指令,根据确定的指定设备的标识和所述待执行指令生成所述宏指令。

7.根据权利要求1所述的系统,其特征在于,所述指令生成设备还包括:

指令分派模块,用于将所述运行指令发送至所述运行设备,以使所述运行设备执行所述运行指令,

其中,所述指令分派模块,包括:

指令汇编子模块,用于根据所述运行指令生成汇编文件;

汇编翻译子模块,用于将所述汇编文件翻译成二进制文件;

指令发送子模块,用于将所述二进制文件发送至所述运行设备,以使所述运行设备根据所述二进制文件执行所述运行指令。

8.根据权利要求1所述的系统,其特征在于,所述运行设备,还包括:

存储模块,所述存储模块包括寄存器、缓存中任意组合,所述缓存包括高速暂存缓存,

所述缓存,用于存储所述数据;

所述寄存器,用于存储所述数据中标量数据。

9.根据权利要求1所述的系统,其特征在于,所述控制模块,包括:

指令存储子模块,用于存储所述运行指令;

指令处理子模块,用于对所述运行指令进行解析,得到所述多个解析指令;

存储队列子模块,用于存储运行指令队列,所述运行指令队列包括所述运行指令和所述多个解析指令,所述运行指令队列所述运行指令和所述多个解析指令按照被执行的先后顺序依次排列。

10.根据权利要求9所述的系统,其特征在于,所述执行模块,包括:

依赖关系处理子模块,用于在确定第一解析指令与所述第一解析指令之前的第零解析指令存在关联关系时,将所述第一解析指令缓存在所述指令存储子模块中,在所述第零解析指令执行完毕后,从所述指令存储子模块中提取所述第一解析指令发送至所述执行模块,

其中,所述第一解析指令与所述第一解析指令之前的第零解析指令存在关联关系包括:

存储所述第一解析指令所需数据的第一存储地址区间与存储所述第零解析指令所需数据的第零存储地址区间具有重叠的区域。

11.根据权利要求1所述的系统,其特征在于,

所述运行设备为CPU、GPU和NPU中的其中一种或任意组合;

所述指令生成设备设置于CPU和/或NPU中;

所述宏指令包括以下指令中的至少一种:计算宏指令、控制宏指令和数据搬运指令,

其中,所述计算宏指令包括神经网络计算宏指令、向量逻辑计算宏指令、矩阵向量计算宏指令、标量计算宏指令和标量逻辑计算宏指令中的至少一种,

所述控制宏指令包括无条件跳转宏指令和有条件跳转宏指令中的至少一种,

数据搬运宏指令包括读宏指令和写宏指令中的至少一种,所述读宏指令包括读神经元

宏指令、读突触宏指令和读标量宏指令中的至少一种，所述写宏指令包括写神经元宏指令、写突触宏指令和写标量宏指令中的至少一种；

所述宏指令包含以下选项中的至少一项：用于执行所述宏指令的指定设备的标识、操作类型、输入地址、输出地址、输入量、输出量、操作数和指令参数，

所述运行指令包含以下选项中的至少一项：所述操作类型、所述输入地址、所述输出地址、所述操作数和所述指令参数。

12.一种机器学习运算装置，其特征在于，所述装置包括：

一个或多个如权利要求1-11任一项所述的神经网络指令处理系统，用于从其他处理装置中获取待运算数据和控制信息，并执行指定的机器学习运算，将执行结果通过I/O接口传递给其他处理装置；

当所述机器学习运算装置包含多个所述神经网络指令处理系统时，所述多个所述神经网络指令处理系统间可以通过特定的结构进行连接并传输数据；

其中，多个所述神经网络指令处理系统通过快速外部设备互连总线PCIE总线进行互联并传输数据，以支持更大规模的机器学习的运算；多个所述神经网络指令处理系统共享同一控制系统或拥有各自的控制系统；多个所述神经网络指令处理系统共享内存或者拥有各自的内存；多个所述神经网络指令处理系统的互联方式是任意互联拓扑。

13.一种组合处理装置，其特征在于，所述组合处理装置包括：

如权利要求12所述的机器学习运算装置、通用互联接口和其他处理装置；

所述机器学习运算装置与所述其他处理装置进行交互，共同完成用户指定的计算操作，

其中，所述组合处理装置还包括：存储装置，该存储装置分别与所述机器学习运算装置和所述其他处理装置连接，用于保存所述机器学习运算装置和所述其他处理装置的数据。

14.一种机器学习芯片，其特征在于，所述机器学习芯片包括：

如权利要求12所述的机器学习运算装置或如权利要求13所述的组合处理装置。

15.一种电子设备，其特征在于，所述电子设备包括：

如权利要求14所述的机器学习芯片。

16.一种板卡，其特征在于，所述板卡包括：存储器件、接口装置和控制器件以及如权利要求14所述的机器学习芯片；

其中，所述机器学习芯片与所述存储器件、所述控制器件以及所述接口装置分别连接；

所述存储器件，用于存储数据；

所述接口装置，用于实现所述机器学习芯片与外部设备之间的数据传输；

所述控制器件，用于对所述机器学习芯片的状态进行监控。

17.一种神经网络指令处理方法，其特征在于，所述方法应用于神经网络指令处理系统，所述系统包括指令生成设备和运行设备，所述方法包括：

通过所述指令生成设备根据接收到的宏指令，确定执行所述宏指令的运行设备，并根据所述宏指令和所述运行设备，生成运行指令；

通过所述运行设备获取数据、神经网络模型以及运行指令，对所述运行指令进行解析，获得多个解析指令，并根据所述数据执行所述多个解析指令，得到执行结果。

18.根据权利要求17所述的方法，其特征在于，所述方法还包括：

通过所述指令生成设备获取备选设备的资源信息，

其中，通过所述指令生成设备根据接收到的宏指令，确定执行所述宏指令的运行设备，包括以下至少一项：

在确定所述宏指令中包含指定设备的标识，且所述指定设备的资源满足执行所述宏指令的执行条件时，将所述指定设备确定为所述运行设备；

在确定所述宏指令中不包含所述指定设备的标识时，根据接收到的宏指令和所述备选设备的资源信息，从所述备选设备中确定出用于执行所述宏指令的运行设备；

在确定所述宏指令中包含所述指定设备的标识，且所述指定设备的资源不满足执行所述宏指令的执行条件时，根据所述宏指令和所述备选设备的资源信息，确定运行设备，

其中，所述执行条件包括：所述指定设备中包含与所述宏指令相对应的指令集，所述资源信息包括所述备选设备所包含的指令集。

19.根据权利要求18所述的方法，其特征在于，所述宏指令包含输入量和输出量中的至少一项，

根据所述宏指令和所述运行设备，生成运行指令，包括：

确定所述宏指令的数据量，根据所述宏指令的数据量、所述宏指令和所述运行设备的资源信息，生成运行指令，

其中，所述数据量是根据所述输入量和所述输出量中的至少一项确定的，所述运行设备的资源信息还包括存储容量、剩余存储容量的至少一项。

20.根据权利要求19所述的方法，其特征在于，根据所述宏指令的数据量、所述宏指令和所述运行设备的资源信息，生成运行指令，包括以下至少一项：

在确定所述运行设备为一个，且所述运行设备的资源不满足执行所述宏指令的容量条件时，根据所述运行设备的运行数据量和所述数据量将所述宏指令拆分成多条运行指令，以使所述运行设备依次执行多条运行指令；

在确定所述运行设备为多个时，根据每个运行设备的运行数据量和所述数据量对所述宏指令进行拆分，生成对应于每个运行设备的运行指令，

其中，运行设备的运行数据量是根据运行设备的资源信息确定的，所述运行指令包含运行输入量和运行输出量中的至少一项，所述运行输入量和所述运行输出量是根据执行所述运行指令的运行设备的运行数据量确定的。

21.根据权利要求17所述的方法，其特征在于，所述方法还包括：

通过所述指令生成设备根据队列排序规则对所述运行指令进行排序，根据排序后的运行指令构建与所述运行设备相对应的指令队列。

22.根据权利要求18所述的方法，其特征在于，所述方法还包括：

通过所述指令生成设备接收待执行指令，根据确定的指定设备的标识和所述待执行指令生成所述宏指令。

23.根据权利要求17所述的方法，其特征在于，所述方法还包括：

通过所述指令生成设备将所述运行指令发送至所述运行设备，以使所述运行设备执行所述运行指令，

其中，通过所述指令生成设备将所述运行指令发送至所述运行设备，以使所述运行设备执行所述运行指令，包括：

根据所述运行指令生成汇编文件；

将所述汇编文件翻译成二进制文件；

将所述二进制文件发送至所述运行设备，以使所述运行设备根据所述二进制文件执行所述运行指令。

24.根据权利要求17所述的方法，其特征在于，所述方法还包括：

通过所述运行设备存储所述数据以及所述数据中的标量数据，

其中，所述运行设备包括存储模块，所述存储模块包括寄存器、缓存中任意组合，所述缓存包括高速暂存缓存，

所述缓存，用于存储所述数据；

所述寄存器，用于存储所述数据中标量数据。

25.根据权利要求17所述的方法，其特征在于，所述方法还包括：

通过所述运行设备存储所述运行指令；

通过所述运行设备对所述运行指令进行解析，得到所述多个解析指令；

通过所述运行设备存储运行指令队列，所述运行指令队列包括所述运行指令和所述多个解析指令，所述运行指令队列所述运行指令和所述多个解析指令按照被执行的先后顺序依次排列。

26.根据权利要求25所述的方法，其特征在于，所述方法还包括：

通过所述运行设备在确定第一解析指令与所述第一解析指令之前的第零解析指令存在关联关系时，缓存所述第一解析指令，在所述第零解析指令执行完毕后，执行缓存的所述第一解析指令，

其中，所述第一解析指令与所述第一解析指令之前的第零解析指令存在关联关系包括：

存储所述第一解析指令所需数据的第一存储地址区间与存储所述第零解析指令所需数据的第零存储地址区间具有重叠的区域。

27.根据权利要求17所述的方法，其特征在于，

所述运行设备为CPU、GPU和NPU中的其中一种或任意组合；

所述指令生成设备设置于CPU和/或NPU中；

所述宏指令包括以下指令中的至少一种：

计算宏指令、控制宏指令和数据搬运指令，

其中，所述计算宏指令包括神经网络计算宏指令、向量逻辑计算宏指令、矩阵向量计算宏指令、标量计算宏指令和标量逻辑计算宏指令中的至少一种，

所述控制宏指令包括无条件跳转宏指令和有条件跳转宏指令中的至少一种，

数据搬运宏指令包括读宏指令和写宏指令中的至少一种，所述读宏指令包括读神经元宏指令、读突触宏指令和读标量宏指令中的至少一种，所述写宏指令包括写神经元宏指令、写突触宏指令和写标量宏指令中的至少一种；

所述宏指令包括以下选项中的至少一项：

用于执行所述宏指令的指定设备的标识、操作类型、输入地址、输出地址、输入量、输出量、操作数和指令参数，

所述运行指令包括以下选项中的至少一项：所述操作类型、所述输入地址、所述输出地

址、所述操作数和所述指令参数。

运算方法、系统及相关产品

技术领域

[0001] 本公开涉及信息处理技术领域，尤其涉及一种神经网络指令处理方法、系统及相关产品。

背景技术

[0002] 随着科技的不断发展，神经网络算法的使用越来越广泛。其在图像识别、语音识别、自然语言处理等领域中都得到了良好的应用。但由于神经网络算法的复杂度越来越高，其模型的规模不断增大。基于图形处理器(Graphics Processing Unit，简称GPU)、中央处理器(Central Processing Unit，简称CPU)的大规模的神经网络模型，要花费大量的计算时间，且耗电量大。相关技术中，对神经网络模型的处理速度进行加快的方式存在无法跨平台处理、处理效率低、开发成本高、易出错等问题。

发明内容

[0003] 有鉴于此，本公开提出了一种神经网络指令处理方法、系统及相关产品，使其能够跨平台使用，提高处理效率，降低出错几率和开发成本。

[0004] 根据本公开的第一方面，提供了一种神经网络指令处理系统，所述系统包括指令生成设备和运行设备，

[0005] 所述指令生成设备，包括：

[0006] 设备确定模块，用于根据接收到的宏指令，确定执行所述宏指令的运行设备；

[0007] 指令生成模块，用于根据所述宏指令和所述运行设备，生成运行指令；

[0008] 所述运行设备，包括：

[0009] 控制模块，用于获取所需数据、神经网络模型以及所述运行指令，对所述运行指令进行解析，获得多个解析指令；

[0010] 执行模块，用于根据所述数据执行所述多个解析指令，得到执行结果。

[0011] 根据本公开的第二方面，提供了一种机器学习运算装置，所述装置包括：

[0012] 一个或多个上述第一方面所述的神经网络指令处理系统，用于从其他处理装置中获取待运算数据和控制信息，并执行指定的机器学习运算，将执行结果通过I/O接口传递给其他处理装置；

[0013] 当所述机器学习运算装置包含多个所述神经网络指令处理系统时，所述多个所述神经网络指令处理系统间可以通过特定的结构进行连接并传输数据；

[0014] 其中，多个所述神经网络指令处理系统通过快速外部设备互连总线PCIE总线进行互联并传输数据，以支持更大规模的机器学习的运算；多个所述神经网络指令处理系统共享同一控制系统或拥有各自的控制系统；多个所述神经网络指令处理系统共享内存或者拥有各自的内存；多个所述神经网络指令处理系统的互联方式是任意互联拓扑。

[0015] 根据本公开的第三方面，提供了一种组合处理装置，所述装置包括：

[0016] 上述第二方面所述的机器学习运算装置、通用互联接口和其他处理装置；

[0017] 所述机器学习运算装置与所述其他处理装置进行交互，共同完成用户指定的计算操作。

[0018] 根据本公开的第四方面，提供了一种机器学习芯片，所述机器学习芯片包括上述第二方面所述的机器学习络运算装置或上述第三方面所述的组合处理装置。

[0019] 根据本公开的第五方面，提供了一种机器学习芯片封装结构，该机器学习芯片封装结构包括上述第四方面所述的机器学习芯片。

[0020] 根据本公开的第六方面，提供了一种板卡，该板卡包括上述第五方面所述的机器学习芯片封装结构。

[0021] 根据本公开的第七方面，提供了一种电子设备，所述电子设备包括上述第四方面所述的机器学习芯片或上述第六方面所述的板卡。

[0022] 根据本公开的第八方面，提供了一种神经网络指令处理方法，所述方法应用于神经网络指令处理系统，所述系统包括指令生成设备和运行设备，所述方法包括：

[0023] 通过所述指令生成设备根据接收到的宏指令，确定执行所述宏指令的运行设备，并根据所述宏指令和所述运行设备，生成运行指令；

[0024] 通过所述运行设备获取数据、神经网络模型以及运行指令，对所述运行指令进行解析，获得多个解析指令，并根据所述数据执行所述多个解析指令，得到执行结果。

[0025] 在一些实施例中，所述电子设备包括数据处理装置、机器人、电脑、打印机、扫描仪、平板电脑、智能终端、手机、行车记录仪、导航仪、传感器、摄像头、服务器、云端服务器、相机、摄像机、投影仪、手表、耳机、移动存储、可穿戴设备、交通工具、家用电器、和/或医疗设备。

[0026] 在一些实施例中，所述交通工具包括飞机、轮船和/或车辆；所述家用电器包括电视、空调、微波炉、冰箱、电饭煲、加湿器、洗衣机、电灯、燃气灶、油烟机；所述医疗设备包括核磁共振仪、B超仪和/或心电图仪。

[0027] 本公开实施例所提供的神经网络指令处理方法、系统及相关产品，该系统包括指令生成设备和运行设备。指令生成设备包括：设备确定模块用于根据接收到的宏指令，确定执行宏指令的运行设备；指令生成模块用于根据宏指令和运行设备，生成运行指令。运行设备包括：控制模块用于获取所需数据、神经网络模型以及运行指令，对运行指令进行解析，获得多个解析指令；执行模块用于根据所述数据执行所述多个解析指令，得到执行结果。该方法、系统及相关产品可跨平台使用，适用性好，指令转换的速度快、处理效率高、出错几率低，且开发的人力、物力成本低。

[0028] 根据下面参考附图对示例性实施例的详细说明，本公开的其它特征及方面将变得清楚。

附图说明

[0029] 包含在说明书中并且构成说明书的一部分的附图与说明书一起示出了本公开的示例性实施例、特征和方面，并且用于解释本公开的原理。

[0030] 图1示出根据本公开一实施例的神经网络指令处理系统的框图。

[0031] 图2示出根据本公开一实施例的神经网络指令处理系统的框图。

[0032] 图3a、图3b示出根据本公开一实施例的神经网络指令处理系统的应用场景的示意

图。

[0033] 图4a、图4b示出根据本公开一实施例的组合处理装置的框图。

[0034] 图5示出根据本公开一实施例的板卡的结构示意图。

[0035] 图6示出根据本公开一实施例的神经网络指令处理方法的流程图。

具体实施方式

[0036] 以下将参考附图详细说明本公开的各种示例性实施例、特征和方面。附图中相同的附图标记表示功能相同或相似的元件。尽管在附图中示出了实施例的各种方面，但是除非特别指出，不必按比例绘制附图。

[0037] 在这里专用的词“示例性”意为“用作例子、实施例或说明性”。这里作为“示例性”所说明的任何实施例不必解释为优于或好于其它实施例。

[0038] 另外，为了更好的说明本公开，在下文的具体实施方式中给出了众多的具体细节。本领域技术人员应当理解，没有某些具体细节，本公开同样可以实施。在一些实例中，对于本领域技术人员熟知的方法、手段、元件和电路未作详细描述，以便于凸显本公开的主旨。

[0039] 图1示出根据本公开一实施例的神经网络指令处理系统的框图。如图1所示，该系统包括指令生成设备100和运行设备200。

[0040] 指令生成设备100包括设备确定模块11和指令生成模块12。设备确定模块11用于根据接收到的宏指令，确定执行宏指令的运行设备。指令生成模块12用于根据宏指令和运行设备，生成运行指令。

[0041] 运行设备200包括控制模块21和执行模块22。控制模块21用于获取所需数据、神经网络模型以及运行指令，对运行指令进行解析，获得多个解析指令。执行模块22用于根据数据执行多个解析指令，得到执行结果。

[0042] 在该实现方式中，宏指令是一种批量处理的称谓，宏指令可以是一种规则或模式，或称语法替换，在遇到宏指令时会自动进行这一规则或模式的替换。宏指令可以是对常用的用于对数据进行计算、控制和搬运等处理的待执行指令整合形成的。

[0043] 在一种可能的实现方式中，宏指令可以包括以下至少一种：计算宏指令、控制宏指令和数据搬运宏指令。其中，计算宏指令可以包括神经网络计算宏指令、向量逻辑计算宏指令、矩阵向量计算宏指令、标量计算宏指令和标量逻辑计算宏指令中的至少一种。控制宏指令可以包括无条件跳转宏指令和有条件跳转宏指令中的至少一种。数据搬运宏指令可以包括读宏指令和写宏指令中的至少一种。读宏指令可以包括读神经元宏指令、读突触宏指令和读标量宏指令中的至少一种。写宏指令可以包括写神经元宏指令、写突触宏指令和写标量宏指令中的至少一种。

[0044] 在一种可能的实现方式中，宏指令可以包含以下选项中的至少一项：用于执行宏指令的指定设备的标识、操作类型、输入地址、输出地址、输入量、输出量、操作数和指令参数。运行指令可以包含以下选项中的至少一项：操作类型、输入地址、输出地址、操作数和指令参数。

[0045] 其中，指定设备的标识可以是指定设备的物理地址、IP地址、名称、编号等标识。标识可以包括数字、字母、符号中的其中一种或任意组合。在宏指令的指定设备的标识的位置为空时，确定该宏指令无指定设备；或者，在宏指令中不包含“指定设备的标识”这个字段

时，确定该宏指令无指定设备。操作类型可以是指该宏指令对数据所进行操作的类型，表征该宏指令的具体类型，如在某宏指令的操作类型为“XXX”时，可以根据“XXX”确定该宏指令对数据所进行的操作的具体类型。根据操作类型可以确定执行该宏指令所需的指令集合，如在某宏指令的操作类型为“XXX”时，其所需的指令集合为进行“XXX”所对应的处理所需的所有指令集。输入地址可以是数据的输入地址、读取地址等获得数据的地址，输出地址可以是被处理后的数据的输出地址、写入地址等存储数据的地址。输入量可以是数据的输入规模、输入长度等表征其数据量大小的信息。输出量可以是数据的输出规模、输出长度等表征其数据量的大小的信息。操作数可以包括寄存器的长度、寄存器的地址、寄存器的标识、立即数等。立即数为在立即寻址方式指令中给出的数。指令参数可以是指对应于该宏指令、与其执行相关的参数。例如，指令参数可以是第二个操作数的地址和长度等。指令参数可以是卷积核的大小、卷积核的步长和卷积核的填充等。

[0046] 在该实现方式中，对于一个宏指令，其必须包括操作码和至少一个操作域，其中操作码即为操作类型，操作域包括指定设备的标识、输入地址、输出地址、输入量、输出量、操作数和指令参数。操作码可以是计算机程序中所规定的要执行操作的那一部分指令或字段(通常用代码表示)，是指令序列号，用来告知执行指令的装置具体需要执行哪一条指令。操作域可以是执行对应的指令所需的所有数据的来源，执行对应的指令所需的所有数据包括参数数据、待运算或待处理的数据、对应的运算方法，或者存储参数数据、待运算或待处理的数据、对应的运算方法的地址等等。

[0047] 应当理解的是，本领域技术人员可以根据需要对宏指令的指令格式以及所包含的内容进行设置，本公开对此不作限制。

[0048] 在本实施例中，设备确定模块11可以根据宏指令确定一个或多个运行设备。指令生成模块12可以生成一个或多个运行指令。在生成的运行指令为多个时，多个运行指令可以在同一个运行设备中被执行，也可以在不同的运行设备中被执行，本公开对此不作限制。

[0049] 本公开实施例所提供的神经网络指令处理系统，该系统包括指令生成设备和运行设备。指令生成设备包括设备确定模块用于根据接收到的宏指令，确定执行宏指令的运行设备；指令生成模块用于根据宏指令和运行设备，生成运行指令。运行设备包括控制模块用于获取所需数据、神经网络模型以及运行指令，对运行指令进行解析，获得多个解析指令；执行模块用于根据所述数据执行所述多个解析指令，得到执行结果。本公开实施例所提供的神经网络指令处理系统，可跨平台使用，适用性好，指令转换的速度快、处理效率高、出错几率低，且开发的人力、物力成本低。

[0050] 图2示出根据本公开一实施例的神经网络指令处理系统的框图。在一种可能的实现方式中，如图2所示，指令生成设备100还可以包括宏指令生成模块13。宏指令生成模块13用于接收待执行指令，根据确定的指定设备的标识和待执行指令生成宏指令。

[0051] 在该实现方式中，指定设备可以是根据待执行指令的操作类型、输入量、输出量等确定的。所接收到的待执行指令可以是一条，也可以是多条。

[0052] 待执行指令可以包括以下至少一种：待执行计算指令、待执行控制指令和待执行数据搬运指令。其中，待执行计算指令可以包括待执行神经网络计算指令、待执行向量逻辑计算指令、待执行矩阵向量计算指令、待执行标量计算指令和待执行标量逻辑计算指令中的至少一种。待执行控制指令可以包括待执行无条件跳转指令和待执行有条件跳转指令中

的至少一种。待执行数据搬运指令可以包括待执行读指令和待执行写指令中的至少一种。待执行读指令可以包括待执行读神经元指令、待执行读突触指令和待执行读标量指令中的至少一种。待执行写指令可以包括待执行写神经元指令、待执行写突触指令和待执行写标量指令中的至少一种。

[0053] 待执行指令可以包含以下选项中的至少一项:操作类型、输入地址、输出地址、输入量、输出量、操作数和指令参数。

[0054] 在该实现方式中,在待执行指令为一个时,可以将确定的指定设备的标识添加到待执行指令中,生成宏指令。举例来说,某待执行指令m为“XXX……param)”。其中,XXX为操作类型,param为指令参数。可以根据该待执行指令m的操作类型“XXX”确定其指定设备m-1。然后,在待执行指令m中添加指定设备m-1的标识(例如,09),生成对应该待执行指令m的宏指令M“XXX 09,……param”。在待执行指令为多个时,可以将确定的每个待执行指令所对应的指定设备的标识添加到待执行指令中,根据带有指定设备的标识的多个待执行指令,生成一个宏指令,或者生成对应的多个宏指令。

[0055] 应当理解的是,本领域技术人员可以根据需要对待执行指令的指令格式以及所包含的内容进行设置,本公开对此不作限制。

[0056] 在一种可能的实现方式中,如图2所示,设备确定模块11可以包括第一确定子模块111。第一确定子模块111用于在确定宏指令中包含指定设备的标识,且指定设备的资源满足执行宏指令的执行条件时,将指定设备确定为运行设备。其中,执行条件可以包括:指定设备中包含与宏指令相对应的指令集。

[0057] 在该实现方式中,宏指令中可以包含执行宏指令的一个或多个指定设备的标识。在宏指令中包含指定设备的标识,且指定设备的资源满足执行条件时,第一确定子模块111可以直接将指定设备确定为运行设备,节省基于宏指令生成运行指令的生成时间,且可以保证所生成的运行指令能够被对应的运行设备所执行。

[0058] 在一种可能的实现方式中,如图2所示,指令生成设备100还可以包括资源获取模块14。设备确定模块11还可以包括第二确定子模块112。资源获取模块14用于获取备选设备的资源信息。第二确定子模块112用于在确定宏指令中不包含指定设备的标识时,根据接收到的宏指令和备选设备的资源信息,从备选设备中确定出用于执行宏指令的运行设备。其中,资源信息可以包括备选设备所包含的指令集。备选设备所包含的指令集可以是对应于一种或多种宏指令的操作类型的指令集合。备选设备所包含的指令集越多,备选设备能够执行的宏指令的类型越多。

[0059] 在该实现方式中,第二确定子模块112在确定宏指令中不包含指定设备的标识时,可以从备选设备中确定出能够执行宏指令的一个或多个运行设备。其中,所确定的运行设备的指令集中包括与宏指令相对应的指令集合。例如,接收到的宏指令为神经网络计算宏指令,可将包含对应于神经网络计算宏指令的指令集的备选设备确定为运行设备,以保证其可以运行生成的运行指令。

[0060] 在一种可能的实现方式中,如图2所示,设备确定模块11还可以包括第三确定子模块113。第三确定子模块113在确定宏指令中包含指定设备的标识,且指定设备的资源不满足执行宏指令的执行条件时,根据宏指令和备选设备的资源信息,确定运行设备。

[0061] 在该实现方式中,第三确定子模块113在确定宏指令中包含指定设备的标识,且指

定设备的资源不满足执行条件时，可以认定该宏指令的指定设备不具备执行宏指令的能力。第三确定子模块113可以从备选设备中确定运行设备，可以将包含与宏指令相对应的指令集的备选设备确定为运行设备。

[0062] 在一种可能的实现方式中，如图2所示，宏指令可以包含输入量和输出量中的至少一项，指令生成模块12还用于确定宏指令的数据量，根据宏指令的数据量、宏指令和运行设备的资源信息，生成运行指令。其中，宏指令的数据量可以是根据输入量和输出量中的至少一项确定的，运行设备的资源信息还可以包括存储容量、剩余存储容量的至少一项。

[0063] 其中，运行设备的存储容量可以是指运行设备的存储器可以容纳的二进制信息量。运行设备的剩余存储容量可以是指去除被占用的存储容量之后，运行设备当前所能用于指令运行的存储容量。运行设备的资源信息能够表征该运行设备的运行能力。存储容量越大、剩余存储容量越大，运行设备的运行能力越强。

[0064] 在该实现方式中，指令生成模块12可以根据每个运行设备的资源信息、宏指令的数据量等，确定拆分宏指令的具体方式，以对宏指令进行拆分，生成与运行设备相对应的运行指令。

[0065] 在一种可能的实现方式中，如图2所示，指令生成模块12可以包括第一指令生成子模块121。第一指令生成子模块121用于在确定运行设备为一个，且在运行设备的资源不满足执行宏指令的容量条件时，根据运行设备的运行数据量和数据量将宏指令拆分成多条运行指令，以使运行设备依次执行多条运行指令。其中，运行设备的运行数据量可以是根据运行设备的资源信息确定的，每条运行指令可以包含运行输入量和运行输出量中的至少一项，运行输入量和运行输出量可以是根据运行数据量确定的。

[0066] 在该实现方式中，运行设备的运行数据量可以是根据运行设备的存储容量或剩余存储容量确定的。容量条件可以是运行设备的运行数据量大于或等于宏指令的数据量，换言之，运行设备的资源不满足执行宏指令的容量条件可以是指：运行设备的运行数据量小于宏指令的数据量。运行输入量和运行输出量需小于或等于运行数据量，以保证所生成的运行指令可以被运行设备执行。多个运行指令中不同运行指令的运行输入量(或运行输出量)可以相同，也可以不同，本公开对此不作限制。

[0067] 在该实现方式中，第一指令生成子模块121在确定运行设备为一个，且在运行设备的资源满足执行宏指令的容量条件时，可以直接将宏指令转化为一个运行指令，还可以将宏指令拆分为多个运行指令，本公开对此不作限制。

[0068] 在一种可能的实现方式中，如图2所示，指令生成模块12可以包括第二指令生成子模块122。第二指令生成子模块122用于在确定运行设备为多个时，根据每个运行设备的运行数据量和数据量对宏指令进行拆分，生成对应于每个运行设备的运行指令。其中，每个运行设备的运行数据量可以是根据每个运行设备的资源信息确定的，运行指令可以包含运行输入量和运行输出量中的至少一项，运行输入量和运行输出量是根据执行运行指令的运行设备的运行数据量确定的。

[0069] 在该实现方式中，运行输入量和运行输出量需小于或等于运行数据量，以保证所生成的运行指令可以运行设备执行。第二指令生成子模块122可以根据每个运行设备的运行数据量，为每个运行设备生成一个或多个运行指令，以供对应的运行设备执行。

[0070] 在上述实现方式中，运行指令中包含运行输入量运行输出量中的至少一项，除了

可以限定运行指令的数据量，使其能够被对应的运行设备执行之外。还可以满足不同运行指令对运行输入量和/或运行输出量的特殊限定需求。

[0071] 在一种可能的实现方式中，对于一些对运行输入量和/或运行输出量没有特殊限定需求的运行指令，其中可以不包含运行输入量和/或运行输出量，可以预先设置默认运行输入量和默认运行输出量，使得运行设备在确定接收到的运行指令中不存在运行输入量、运行输出量时，可以将默认运行输入量、默认运行输出量作为该运行指令的运行输入量、运行输出量。通过预设默认运行输入量和默认运行输出量的方式，可以简化运行指令的生成过程，节省运行指令的生成时间。

[0072] 在一种可能的实现方式中，可以将预先设置针对不同类型的宏指令的默认输入量和默认输出量。在宏指令中不包含输入量和输出量时，可以将对应的预先设置的默认输入量和默认输出量作为宏指令的输入量和输出量。进而根据默认输入量和/或默认输出量确定宏指令的数据量，并根据宏指令的数据量、宏指令和运行设备的资源信息，生成运行指令。在宏指令中不包含输入量和输出量时，所生成的运行指令可以不包含运行输入量和运行输出量，也可以包含运行输入量和运行输出量中的至少一项。在运行指令中不包含运行输入量和/或运行输出量时，运行设备可以根据预先设置的默认运行输入量和/或默认运行输出量执行运行指令。

[0073] 在一种可能的实现方式中，指令生成模块12还可以根据宏指令以及预先设置的宏指令拆分规则，对宏指令进行拆分生成运行指令。宏指令拆分规则可以是根据常规的宏指令拆分方式(例如，根据宏指令的处理过程等进行拆分)，结合所有备选设备能够执行的指令的运行数据量阈值确定的。将宏指令拆分成运行输入量以及运行输出量均小于或等于运行数据量阈值的运行指令，以保证生成的运行指令可以在其对应的运行设备(运行设备为备选设备中的任意一个)中被执行。其中，可以比较所有备选设备的存储容量(或剩余存储容量)，将确定出的最小的存储容量(或剩余存储容量)确定为所有备选设备能够执行的指令的运行数据量阈值。

[0074] 应当理解的是，本领域技术人员可以根据实际需要对运行指令的生成方式进行设置，本公开对此不作限制。

[0075] 在本实施例中，指令生成模块根据宏指令所生成的运行指令可以是待执行指令，也可以是对待执行指令进行解析所获得的解析后的一个或多个指令，本公开对此不作限制。

[0076] 在一种可能的实现方式中，如图2所示，指令生成设备100还可以包括队列构建模块15。队列构建模块15用于根据队列排序规则对运行指令进行排序，根据排序后的运行指令构建与运行设备相对应的指令队列。

[0077] 在该实现方式中，可以为每个运行设备构建与之唯一对应的指令队列。可以按照指令队列中运行指令的排序，依次向指令队列唯一对应的运行设备发送运行指令；或者可以将指令队列发送至运行设备，以使运行设备按照指令队列中运行指令的排序依次执行其中的运行指令。通过上述方式，可保证运行设备按照指令队列执行运行指令，避免运行指令被错误、延误执行，避免运行指令被遗漏执行。

[0078] 在该实现方式中，队列排序规则可以是根据执行运行指令的预计执行时长、运行指令的生成时间、与运行指令自身相关的运行输入量、运行输出量、操作类型等信息确定

的,本公开对此不作限制。

[0079] 在一种可能的实现方式中,如图2所示,指令生成设备100还可以包括指令分派模块16。指令分派模块16用于将运行指令发送至运行设备,以使运行设备执行运行指令。

[0080] 在该实现方式中,在运行设备所执行的运行指令为一个时,可以直接将该运行指令发送至运行设备。在运行设备所执行的运行指令为多个时,可以将多个运行指令全部发送至运行设备,以使运行设备依次执行多个运行指令。还可以将多个运行指令依次发送给与之对应的运行设备,其中,每次在运行设备执行完成当前运行指令之后,向运行设备发送与之对应的下一个运行指令。本领域技术人员可以对向运行设备发送运行指令的方式进行设置,本公开对此不作限制。

[0081] 在一种可能的实现方式中,如图2所示,指令分派模块16可以包括指令汇编子模块161、汇编翻译子模块162和指令发送子模块163。指令汇编子模块161用于根据所述运行指令生成汇编文件。汇编翻译子模块162用于将汇编文件翻译成二进制文件。指令发送子模块163用于将二进制文件发送至运行设备,以使运行设备根据二进制文件执行运行指令。

[0082] 通过上述方式,可以降低运行指令的数据量,节省向运行设备发送运行指令的时间,提高宏指令的转换、执行速度。

[0083] 在该实现方式中,在二进制文件被发送至运行设备之后,运行设备可以对接收到的二进制文件进行译码获得对应的运行指令,并执行所获得的运行指令,获得执行结果。

[0084] 在一种可能的实现方式中,运行设备可以为CPU、GPU和嵌入式神经网络处理器(Neural-network Processing Unit,简称NPU)中的其中一种或任意组合。这样,提高了指令生成设备根据宏指令生成运行指令的速度。

[0085] 在一种可能的实现方式中,该指令生成设备100可以设置于CPU和/或NPU中。以实现通过CPU和/或NPU实现根据宏指令生成运行指令的过程,为指令生成设备的实现提供了更多的可能方式。

[0086] 在一种可能的实现方式中,运行设备200还包括存储模块23。该存储模块23可以包括寄存器和缓存中的至少一种,缓存可以包括高速暂存缓存。缓存可以用于存储数据。寄存器可以用于存储数据中的标量数据。

[0087] 在一种可能的实现方式中,控制模块21可以包括指令存储子模块211和指令处理子模块212。指令存储子模块211用于存储运行指令。指令处理子模块212用于对运行指令进行解析,得到多个解析指令。

[0088] 在一种可能的实现方式中,控制模块21还可以包括存储队列子模块213。存储队列子模块213用于存储运行指令队列,该运行指令队列中包含运行设备所需执行的运行指令以及多个解析指令。在运行指令队列中所有指令按照执行的先后顺序依次排列。

[0089] 在一种可能的实现方式中,执行模块22还可以包括依赖关系处理子模块221。依赖关系处理子模块221用于在确定第一解析指令与第一解析指令之前的第零解析指令存在关联关系时,将第一解析指令缓存在指令存储子模块中,在第零解析指令执行完毕后,从指令存储子模块中提取第一解析指令发送至执行模块。

[0090] 其中,第一解析指令与第一解析指令之前的第零解析指令存在关联关系可以包括:存储第一解析指令所需数据的第一存储地址区间与存储第零解析指令所需数据的第零存储地址区间具有重叠区域。反之,第一解析指令与第零解析指令之间没有关联关系可以

是第一存储地址区间与第零存储地址区间没有重叠区域。

[0091] 在一种可能的实现方式中，计算宏指令是指用于进行数据计算的宏指令，数据计算可以包括机器学习计算、神经网络计算、向量逻辑计算、矩阵向量计算、标量计算和标量逻辑计算中的至少一种。

[0092] 神经网络计算宏指令可以是指用于对神经网络算法进行计算的宏指令。例如，对卷积运算(convolutional computation)、池化运算(Pooling)等神经网络算法进行计算的卷积计算宏指令、池化计算宏指令等。不同类型的神经网络计算宏指令对应于不同的操作类型。例如，卷积计算宏指令所对应的操作类型可以为CONV。

[0093] 向量逻辑计算宏指令可以是指用于对向量进行逻辑运算的宏指令。例如，对向量进行“与”、“比较”、“或”等逻辑计算的宏指令。不同类型的向量逻辑计算宏指令对应于不同的操作类型。例如，向量与计算宏指令所对应的操作类型可以为VAND、向量或计算宏指令所对应的操作类型可以为VOR。

[0094] 矩阵向量计算宏指令可以是指用于对矩阵和向量进行计算的宏指令。例如，对矩阵和向量进行矩阵乘向量计算、向量乘矩阵计算、张量计算、矩阵相加计算、矩阵相减计算等计算的宏指令。不同类型的矩阵向量计算宏指令对应于不同的操作类型，例如，矩阵相加计算宏指令所对应的操作类型为MADD，矩阵乘向量计算宏指令所对应的操作类型为MMV。

[0095] 标量计算宏指令可以是指用于对标量进行算术运算的宏指令。例如，对标量进行标量相加、标量相减、标量相乘、标量相除等计算的宏指令。不同类型的标量计算宏指令对应于不同的操作类型。例如，标量相减宏指令所对应的操作类型为SSUB、标量相加宏指令所对应的操作类型为SADD。

[0096] 标量逻辑计算宏指令可以是用于对标量进行逻辑运算的宏指令。例如，对标量进行“与”、“比较”、“或”、“非”等逻辑运算的宏指令。不同类型的标量逻辑计算宏指令对应于不同的操作类型，例如，标量与计算宏指令所对应的操作类型可以为SAND、标量或计算宏指令所对应的操作类型可以为SOR。

[0097] 在一种可能的实现方式中，控制宏指令是指用于控制指令流跳转至目标跳转位置的宏指令。无条件跳转宏指令可以是用于对指令流进行控制，使其无条件的跳转至指定位置的宏指令。有条件跳转宏指令可以是用于对指令流进行控制，使有条件跳转宏指令在其需要满足的条件为真时，跳转至指定位置的宏指令。

[0098] 在一种可能的实现方式中，数据搬运宏指令是指用于对数据进行读入、写入等搬运处理的宏指令。读宏指令可以是指将数据从内存中读入到存储数据的位置的宏指令，可以是指用于进行数据读入的宏指令。根据数据类型的不同，读宏指令可以包括用于读入神经元数据的读神经元宏指令、用于读入突触数据的读突触宏指令和用于读入标量数据的读标量宏指令。写宏指令可以是指将数据从其存储位置写入到内存中的宏指令，可以是指用于进行数据写入的宏指令。根据数据类型的不同，写宏指令可以包括用于写入神经元数据的写神经元宏指令、用于写入突触数据的写突触宏指令和用于写入标量数据的写标量宏指令。其中，神经元数据即为神经网络算法中的输入神经元、输出神经元，突触数据即为神经网络算法中的权值。

[0099] 需要说明的是，尽管以上述实施例作为示例介绍了神经网络指令处理系统如上，但本领域技术人员能够理解，本公开应不限于此。事实上，用户完全可根据个人喜好和/或

实际应用场景灵活设定各模块，只要符合本公开的技术方案即可。

[0100] 应用示例

[0101] 以下结合“神经网络指令处理系统的工作过程”作为一个示例性应用场景，给出根据本公开实施例的应用示例，以便于理解神经网络指令处理系统的流程。本领域技术人员应理解，以下应用示例仅仅是出于便于理解本公开实施例的目的，不应视为对本公开实施例的限制。

[0102] 首先，对宏指令的指令格式、待执行指令的指令格式及运行设备执行运行指令的过程进行描述，以下是为具体示例。

[0103] 宏指令的指令格式可以为如下格式示例。

[0104] 神经网络计算宏指令的指令格式可以是：

[0105] Type device_id,input_addr,output_addr,input_h,input_w,input_c,output_h,output_w,output_c,[param1,param2,…]

[0106] 其中，Type为操作类型，device_id为指定设备的标识，input_addr为输入地址，output_addr为输出地址，input_h、input_w、input_c为输入的神经元规模(即输入量)，output_h、output_w、output_c为输出的神经元规模(即输出量)，param1、param2为指令参数。

[0107] 对于神经网络计算宏指令，其必须包含操作类型、输入地址和输出地址，且根据神经网络计算宏指令所生成的运行指令也须包含操作类型、运行输入地址和运行输出地址，运行输入地址和运行输出地址是分别根据输入地址、输出地址确定的。

[0108] 以卷积计算宏指令为例，其指令格式为:CONV device_id,input_addr,output_addr,input_h,input_w,input_c,output_h,output_w,output_c,kernel,stride,pad。调用时卷积计算宏指令可以为如下示例：

[0109] @CONV#0,#4,#500,#5,#5,#32,#3,#3,#16,#3,#1,#0

[0110] 其中，该卷积计算宏指令的操作类型为CONV。指定设备为设备0。数据的输入地址为地址4。数据的输出地址为地址500。数据的输入量为5x5x32。数据的输出量为3x3x16。卷积核的大小为3，卷积核的步长为1，卷积核的填充为0。

[0111] 若根据上述卷积计算宏指令示例生成的运行指令为“@CONV#4,#500,#5,#5,#32,#3,#3,#16,#3,#1,#0”为例。运行设备在接收到该运行指令之后，其执行过程为:从地址4处获取到输入量为5x5x32的数据。按照该运行指令中卷积核的大小(3)、步长为(1)和填充(0)对输入量为5x5x32的数据进行卷积运算，得到数据量为3x3x16的执行结果时，将该执行结果存储至输出地址500。

[0112] 向量逻辑计算宏指令的指令格式可以是：

[0113] Type device_id,input_addr,output_addr,input_size,output_size,[param1,param2,…]

[0114] 其中，Type为操作类型，device_id为指定设备的标识，input_addr为输入地址，output_addr为输出地址，input_size为输入向量的大小(即输入量)，output_size为输出向量的大小(即输出量)，param1、param2为指令参数。指令参数可以是第二个操作数的地址和长度。

[0115] 对于向量逻辑计算宏指令，其必须包含操作类型、输入地址和输出地址，且根据向

量逻辑计算宏指令所生成的运行指令也须包含操作类型、运行输入地址和运行输出地址，运行输入地址和运行输出地址是分别根据输入地址、输出地址确定的。

[0116] 以根据某个向量逻辑计算宏指令所生成的运行指令为“@VAND#501,#7,#33,#4”为例。运行设备在接收到该运行指令后，其执行过程为：从输入地址501处获取到大小为33的输入向量，对该输入向量进行“与”逻辑运算，获得大小为4的输出向量，并将该大小为4的输出向量作为执行结果存储至输出地址7处。

[0117] 矩阵向量计算宏指令的指令格式可以是：

[0118] Type device_id,input_addr,output_addr,input_size,output_size,[param1,param2,…]

[0119] 其中，Type为操作类型，idevice_id为指定设备的标识，input_addr为输入地址，output_addr为输出地址，input_size为输入向量的大小(即输入量)，output_size为输出向量的大小(即输出量)，param1、param2为指令参数。指令参数可以是第二个操作数的地址和长度。

[0120] 对于矩阵向量计算宏指令，其必须包含操作类型、输入地址和输出地址，且根据矩阵向量计算宏指令所生成的运行指令也须包含操作类型、运行输入地址和运行输出地址，运行输入地址和运行输出地址是分别根据输入地址、输出地址确定的。

[0121] 以根据某个矩阵向量计算宏指令生成的运行指令为“@MADD#502,#8,#34,#5”为例。运行设备在接收到该运行指令后，其执行过程为：从输入地址502处获得大小为34的输入矩阵向量。对该输入矩阵向量进行矩阵相加计算，获得大小为5的输出矩阵向量。将该大小为5的输出矩阵向量作为执行结果存储至存储地址8处。

[0122] 标量计算宏指令的指令格式可以是：

[0123] Type device_id,op1,op2,ans

[0124] 其中，Type为操作类型，device_id为指定设备的标识，op1、op2为两个操作数。Ans为标量计算宏指令计算结果的存放地址或者用于存放计算结果的寄存器的标识。其中，所获取的标量的大小、对获取的标量进行计算所输出的标量的大小可以预先设定。

[0125] 对于标量计算宏指令，其必须包含操作类型、第一操作数、第二操作数和输出地址。且根据标量计算宏指令所生成的运行指令也须包含操作类型、第一运行操作数、第二运行操作数和运行输出地址。其中，第一运行操作数、第二运行操作数和运行输出地址是分别根据第一操作数、第二操作数和输出地址确定的。

[0126] 以根据某个标量计算宏指令生成的运行指令为“@SADD#503,#504,#3”为例。运行设备在接收到该运行指令后，其执行过程为：从寄存器的地址503中获取第一标量以及从寄存器的地址504中获取第二标量，将第一标量和第二标量相加，并将相加计算所获得的结果作为执行结果存储至寄存器的存储地址3处。

[0127] 标量逻辑计算宏指令的指令格式可以是：

[0128] Type device_id,op1,op2,ans

[0129] 其中，Type为操作类型，device_id为指定设备的标识，op1、op2为两个操作数。Ans为标量逻辑计算宏指令计算结果的存放地址或者用于存放计算结果的寄存器的标识。其中，所获取的标量的大小、对获取的标量进行计算所输出的标量的大小可以预先设定。

[0130] 对于标量逻辑计算宏指令，其必须包含操作类型、第一操作数、第二操作数和输出

地址。且根据标量逻辑计算宏指令所生成的运行指令也须包含操作类型、第一运行操作数、第二运行操作数和运行输出地址。其中，第一运行操作数、第二运行操作数和运行输出地址是分别根据第一操作数、第二操作数和输出地址确定的。

[0131] 以根据某个标量逻辑计算宏指令生成的运行指令为“@SAND#703,#704,#8”为例。运行设备在接收到该运行指令后，其执行过程为：从寄存器的地址703中获取第一标量以及从寄存器的地址704中获取第二标量，对第一标量和第二标量进行“与”逻辑运算，并将所获得的结果作为执行结果存储至寄存器的存储地址8处。

[0132] 无条件跳转宏指令的指令格式可以是：

[0133] Jump device_id,src

[0134] 其中，Jump为无条件跳转宏指令所对应的操作类型，device_id为指定设备的标识，src为指令流所需跳转到的目标跳转位置。目标跳转位置可以是寄存器的长度、寄存器的地址、寄存器的标识、立即数等。

[0135] 对于无条件跳转宏指令，其必须包含操作类型和目标跳转位置，且根据无条件跳转宏指令所生成的运行指令也须包含操作类型和运行目标跳转位置。其中，运行目标跳转位置是根据目标跳转位置确定的。

[0136] 以根据某个无条件跳转宏指令所生成的运行指令为“@Jump#505”为例。运行设备在接收到该运行指令后，其执行过程为：将当前指令流跳转至地址505处继续执行。

[0137] 有条件跳转宏指令的指令格式可以是：

[0138] CB device_id,src,condition

[0139] 其中，CB为有条件跳转宏指令所对应的操作类型，device_id为指定设备的标识，src为指令流所需跳转到的目标跳转位置，condition为跳转的条件。例如，condition可以为“寄存器的值为零是否为真”，在寄存器的值为零时，可以跳转至目标跳转位置。目标跳转位置可以是寄存器的长度、寄存器的地址、寄存器的标识、立即数等。

[0140] 对于有条件跳转宏指令，其必须包含操作类型和目标跳转位置，且根据有条件跳转宏指令所生成的运行指令也须包含操作类型和运行目标跳转位置。其中，运行目标跳转位置是根据目标跳转位置确定的。

[0141] 以根据某个有条件跳转宏指令所生成的运行指令为“@CB#506#h”为例。运行设备在接收到该运行指令后，其执行过程为：判断其跳转条件“h”是否为真，在“h”为真时，将当前指令流跳转至地址506处继续执行。

[0142] 读神经元宏指令的指令格式可以是：

[0143] NLOAD device_id,src_addr,des_addr,size

[0144] 其中，NLOAD为读神经元宏指令所对应的操作类型，device_id为指定设备的标识，src_addr为读取神经元数据的数据读入地址，des_addr为存储读取神经元数据所需的加密方式的数据加密方式地址，size为神经元数据的读入量。

[0145] 以根据某个读神经元宏指令所生成的运行指令为“@NLOAD#505#506#9”为例。运行设备在接收到该运行指令后，其执行过程为：从地址506处获取到神经元数据的加密方式，根据该加密方式从地址505处读取读入量为9的神经元数据。

[0146] 读突触宏指令的指令格式可以是：

[0147] WLOAD device_id,src_addr,des_addr,size

[0148] 其中,WLOAD为读突触宏指令所对应的操作类型,device_id为指定设备的标识,src_addr为读取突触数据的数据读入地址,des_addr为存储读取突触数据所需的加密方式的数据加密方式地址,size为突触数据的读入量。

[0149] 以根据某个读突触宏指令所生成的运行指令为“@WLOAD#507#508#10”为例。运行设备在接收到该运行指令后,其执行过程为:从地址508处获取到读入突触数据的加密方式,根据该加密方式从地址507处读取读入量为10的突触数据。

[0150] 读标量宏指令的指令格式可以是:

[0151] SLOAD device_id,src,des

[0152] 其中,SLOAD为读标量宏指令所对应的操作类型,device_id为指定设备的标识,src为读取标量的数据读入地址,des为存储读取标量数据所需的加密方式的数据加密方式地址。

[0153] 以根据某个读标量宏指令所生成的运行指令为“@SLOAD#601#602”为例。运行设备在接收到该运行指令后,其执行过程为:从地址602处获取到标量数据的加密方式,根据该加密方式从地址601处读取其存储的标量数据。

[0154] 其中,读神经元宏指令、读突触宏指令和读标量宏指令中所包含的数据读入地址以及数据加密方式地址,可以是寄存器的地址、编号、名称等标识。对于读神经元宏指令、读突触宏指令和读标量宏指令,其必须包含操作类型、数据读入地址、数据加密方式地址,运行指令中也须包含操作类型、运行数据读入地址和运行数据加密方式地址。其中,运行数据读入地址和运行数据加密方式地址分别是根据数据读入地址和数据加密方式地址确定的。

[0155] 写神经元宏指令的指令格式可以是:

[0156] NSTORE device_id,src_addr,des_addr,size

[0157] 其中,NSTORE为写神经元宏指令所对应的操作类型,device_id为指定设备的标识,src_addr为写入神经元数据的数据写入地址,des_addr为存储写入神经元数据所需的加密方式的数据加密方式地址,size为数据的写入量。

[0158] 以根据某个写神经元宏指令所生成的运行指令为“@NSTORE#603#604#14”为例。运行设备在接收到该运行指令后,其执行过程为:从地址604处获取到神经元数据的加密方式,根据该加密方式将待写入的神经元数据写入地址14处。

[0159] 写突触宏指令的指令格式可以是:

[0160] WSTORE device_id,src_addr,des_addr,size

[0161] 其中,WSTORE为写突触宏指令所对应的操作类型,device_id为指定设备的标识,src_addr为写入突触数据的数据写入地址,des_addr为存储写入突触数据所需的加密方式的数据加密方式地址,size为数据的写入量。

[0162] 以根据某个写突触宏指令所生成的运行指令为“@WSTORE#605#606#15”为例。运行设备在接收到该运行指令后,其执行过程为:从地址606处获取到突触数据的加密方式,根据该加密方式将待写入的突触数据写入地址14处。

[0163] 写标量宏指令的指令格式可以是:

[0164] SSTORE device_id,src,des

[0165] 其中,SSTORE为写标量宏指令所对应的操作类型,device_id为指定设备的标识,src为写入标量数据的数据写入地址,des为存储写入标量数据所需的加密方式的数据加密

方式地址。

[0166] 以根据某个写标量宏指令所生成的运行指令为“@SSTORE#607#608”为例。运行设备在接收到该运行指令后，其执行过程为：从地址608处获取到标量数据的加密方式，根据该加密方式将待写入的标量数据写入地址14处。

[0167] 其中，写神经元宏指令、写突触宏指令和写标量宏指令中所包含的数据写入地址和数据加密方式地址可以是寄存器的地址、编号、名称等标识。对于写神经元宏指令、写突触宏指令和写标量宏指令，其必须包含操作类型、数据写入地址、数据加密方式地址，运行指令中也须包含操作类型、运行数据写入地址和运行数据加密方式地址。其中，运行数据写入地址和运行数据加密方式地址分别是根据数据写入地址和数据加密方式地址确定的。

[0168] 待执行指令的指令格式可以为如下格式示例。

[0169] 待执行神经网络计算指令的指令格式可以是：

[0170] Type input_addr,output_addr,input_h,input_w,input_c,output_h,output_w,output_c,[param1,param2,…]

[0171] 其中，Type为操作类型，input_addr为输入地址，output_addr为输出地址，input_h、input_w、input_c为输入的神经元规模(即输入量)，output_h、output_w、output_c为输出的神经元规模(即输出量)，param1、param2为指令参数。

[0172] 以待执行卷积指令为例，其指令格式为CONV input_addr,output_addr,input_h,input_w,input_c,output_h,output_w,output_c,kernel,stride,pad。调用时待执行卷积指令可以为：

[0173] @CONV#6,#500,#5,#5,#32,#3,#3,#16,#3,#1,#0

[0174] 其中，该待执行卷积指令的操作类型为卷积神经网络计算。数据的输入地址为地址6。数据的输出地址为地址500。数据的输入量为5x5x32。数据的输出量为3x3x16。卷积核的大小为3，卷积核的步长为1，卷积核的填充为0。

[0175] 待执行向量逻辑计算指令的指令格式可以是：

[0176] Type input_addr,output_addr,input_size,output_size,[param1,param2,…]

[0177] 其中，Type为操作类型，input_addr为输入地址，output_addr为输出地址，input_size为输入向量的大小(即输入量)，output_size为输出向量的大小(即输出量)，param1、param2为指令参数。指令参数可以是第二个操作数的地址和长度。

[0178] 待执行矩阵向量计算指令的指令格式可以是：

[0179] Type input_addr,output_addr,input_size,output_size,[param1,param2,…]

[0180] 其中，Type为操作类型，input_addr为输入地址，output_addr为输出地址，input_size为输入向量的大小(即输入量)，output_size为输出向量的大小(即输出量)，param1、param2为指令参数。指令参数可以是第二个操作数的地址和长度。

[0181] 待执行标量计算指令的指令格式可以是：

[0182] Type op1,op2,ans

[0183] 其中，Type为操作类型，op1、op2为两个操作数。Ans为待执行标量计算指令计算结果的存放地址或者用于存放计算结果的寄存器的标识。

[0184] 待执行标量逻辑计算指令的指令格式可以是：

[0185] Type op1,op2,ans

[0186] 其中,Type为操作类型,op1、op2为两个操作数。Ans为待执行标量逻辑计算指令计算结果的存放地址或者用于存放计算结果的寄存器的标识。

[0187] 待执行无条件跳转指令的指令格式可以是:

[0188] Jump src

[0189] 其中,Jump为待执行无条件跳转指令所对应的操作类型,src为指令流所需跳转到的目标跳转位置。

[0190] 待执行有条件跳转指令的指令格式可以是:

[0191] CB src,condition

[0192] 其中,CB为待执行有条件跳转指令所对应的操作类型,src为指令流所需跳转到的目标跳转位置,condition为跳转的条件。例如,condition可以为“寄存器的值为零是否为真”,在寄存器的值为零时,可以跳转至目标跳转位置。

[0193] 待执行读神经元指令的指令格式可以是:

[0194] NLOAD src_addr,des_addr,size

[0195] 其中,NLOAD为待执行读神经元指令所对应的操作类型,src_addr为读取神经元数据数据的数据读入地址,des_addr为存储读取神经元数据所需的加密方式的数据加密方式地址,size为神经元数据的读入量。

[0196] 待执行读突触指令的指令格式可以是:

[0197] WLOAD src_addr,des_addr,size

[0198] 其中,WLOAD为待执行读突触指令所对应的操作类型,src_addr为读取突触数据的数据读入地址,des_addr为存储读取突触数据所需的加密方式的数据加密方式地址,size为突触数据的读入量。

[0199] 待执行读标量指令的指令格式可以是:

[0200] SLOAD src,des

[0201] 其中,SLOAD为待执行读标量指令所对应的操作类型,src为读取标量数据的数据读入地址,des为存储读取标量数据所需的加密方式的数据加密方式地址。

[0202] 待执行写神经元指令的指令格式可以是:

[0203] NSTORE src_addr,des_addr,size

[0204] 其中,NSTORE为待执行写神经元指令所对应的操作类型,src_addr为写入神经元数据的数据写入地址,des_addr为存储写入神经元数据所需的加密方式的数据加密方式地址,size为神经元数据的写入量。

[0205] 待执行写突触指令的指令格式可以是:

[0206] WSTORE src_addr,des_addr,size

[0207] 其中,WSTORE为待执行写突触指令所对应的操作类型,src_addr为写入突触数据的数据写入地址,des_addr为存储写入突触数据所需的加密方式的数据加密方式地址,size为突触数据的写入量。

[0208] 待执行写标量原指令的指令格式可以是:

[0209] SSTORE src,des

[0210] 其中,SSTORE为待执行写标量指令所对应的操作类型,src为写入标量数据的数据写入地址,des为存储写入标量数据所需的加密方式的数据加密方式地址。

[0211] 图3a、图3b示出根据本公开一实施例的神经网络指令处理系统的应用场景的示意图。神经网络指令处理系统可以包括上述备选设备和指令生成设备。如图3a、图3b所示，用于执行宏指令的备选设备可以为多个，备选设备可以为CPU-1、CPU-2、…、CPU-n，NPU-1、NPU-2、…、NPU-n和GPU-1、GPU-2、…、GPU-n，备选设备被选定用于执行对应的运行指令即为运行设备。计算指令处理系统根据某计算宏指令生成并执行运行指令的工作过程及原理如下。

[0212] 资源获取模块14

[0213] 获取备选设备的资源信息，该资源信息包括备选设备的剩余存储容量、存储容量和备选设备所包含的指令集。资源获取模块14将获取到的备选设备的资源信息发送至设备确定模块11和指令生成模块12。

[0214] 设备确定模块11(包括第一确定子模块111、第二确定子模块112和第三确定子模块113)

[0215] 在接收到宏指令时，根据接收到的宏指令，确定执行宏指令的运行设备。例如，接收到如下宏指令。其中，宏指令可以是来自不同的平台的。

[0216] 宏指令1:@XXX#01……

[0217] 宏指令2:@SSS#02……

[0218] 宏指令3:@DDD#04……

[0219] 宏指令4:@NNN……

[0220] 第一确定子模块111在确定在宏指令中包含指定设备的标识，且确定该指定设备中包含与宏指令相对应的指令集时，第一确定子模块111可以将该指定设备确定为执行宏指令的运行设备，并将确定的运行设备的标识发送至指令生成模块12。例如，第一确定子模块111可以将标识01所对应的指定设备如CPU-2(CPU-2中包含与宏指令1相对应的指令集)确定为用于执行宏指令1的运行设备。可以将标识02所对应的指定设备如CPU-1(CPU-1中包含与宏指令2相对应的指令集)确定为用于执行宏指令2的运行设备。

[0221] 第三确定子模块113在确定在宏指令中包含指定设备的标识，且确定该指定设备中不包含与宏指令相对应的指令集时，第三确定子模块113可以将包含与宏指令相对应的指令集的备选设备确定为运行设备，并将确定的运行设备的标识发送至指令生成模块12。例如，第三确定子模块113在确定标识04所对应的指定设备中不包含与宏指令3相对应的指令集时，可以将包含与宏指令3的操作类型DDD相对应的指令集的备选设备如NPU-n、NPU-2，确定为用于执行宏指令3的运行设备。

[0222] 第二确定子模块112在确定宏指令中不存在指定设备的标识(指定设备的标识所对应的位置为空，或者在宏指令中不包含“指定设备的标识”这个字段)时，第二确定子模块112可以根据该宏指令和备选设备的资源信息，从备选设备中确定出运行设备(具体确定过程详见上文第二确定子模块112的相关描述)，并将确定出的运行设备的标识发送至指令生成模块12。例如，由于宏指令4中不存在指定设备的标识，第二确定子模块112可以根据宏指令4的操作类型NNN和备选设备的资源信息(所包含的指令集)，从备选设备中确定出用于执行宏指令4的运行设备，例如，GPU-n(GPU-n中包含与操作类型NNN相对应的指令集)。

[0223] 指令生成模块12(包括第一指令生成模块121和第二指令生成模块122)

[0224] 第一指令生成模块121在运行设备为一个，且运行设备的资源不满足执行宏指令

的容量条件时，根据运行设备的运行数据量和数据量将宏指令拆分成多条运行指令，并将多条运行指令发送至队列构建模块15。例如，根据宏指令2的数据量和运行设备CPU-1的运行数据量生成多条运行指令2-1、2-2、…、2-n。根据宏指令4的数据量和运行设备GPU-n的运行数据量生成多条运行指令4-1、4-2、…、4-n。

[0225] 第一指令生成模块121在确定运行设备为一个，且运行设备的资源满足执行宏指令的容量条件时，可以根据宏指令生成一条运行指令，并将其发送至队列构建模块15。例如，根据宏指令1的数据量和运行设备CPU-2的运行数据量生成一条运行指令1-1。

[0226] 第二指令生成模块122在确定运行设备为多个时，根据每个运行设备的运行数据量和宏指令的数据量对宏指令进行拆分，生成对应于每个运行设备的运行指令，并将其发送至队列构建模块15。例如，根据宏指令3的数据量和运行设备NPU-n的运行数据量、运行设备NPU-2的运行数据量，为运行设备NPU-n生成多条运行指令3-1、3-2、…、3-n，为运行设备NPU-2生成多条运行指令3’-1、3’-2、…、3’-n。

[0227] 队列构建模块15

[0228] 在接收到运行指令时，根据队列排序规则对每个运行设备所需执行的所有运行指令进行排序，根据排序后的运行指令为每个运行设备构建与之唯一对应的指令队列，并将指令队列发送至指令分派模块16。具体地，

[0229] 对于被运行设备CPU-2执行的一条运行指令1-1。所构建的对应于运行设备CPU-2的指令队列CPU-2”仅包括运行指令1-1。

[0230] 对于被运行设备CPU-1执行的多条运行指令2-1、2-2、…、2-n。根据队列排序规则对多条运行指令2-1、2-2、…、2-n进行排序，根据排序后的多条运行指令2-1、2-2、…、2-n构建与运行设备CPU-1相对应的指令队列CPU-1”。

[0231] 对于被运行设备NPU-n执行的多条运行指令3-1、3-2、…、3-n。根据队列排序规则对多条运行指令3-1、3-2、…、3-n进行排序，根据排序后的多条运行指令3-n、…、3-2、3-1构建与运行设备NPU-n相对应的指令队列NPU-n”。

[0232] 对于被运行设备NPU-2执行的多条运行指令3’-1、3’-2、…、3’-n。根据队列排序规则对多条运行指令3’-1、3’-2、…、3’-n进行排序，根据排序后的多条运行指令3’-n、…、3’-2、3’-1构建与运行设备NPU-2相对应的指令队列NPU-2”。

[0233] 对于被运行设备GPU-n执行的多条运行指令4-1、4-2、…、4-n。根据队列排序规则对多条运行指令4-1、4-2、…、4-n进行排序，根据排序后的多条运行指令4-1、4-2、…、4-n构建与运行设备GPU-n相对应的指令队列GPU-n”。

[0234] 指令分派模块16

[0235] 在接收到指令队列之后，将每个指令队列中的运行指令，依次发送至对应的运行设备中，以使运行设备执行运行指令。例如，将指令队列CPU-2”中包括的运行指令1-1发送至其对应的运行设备CPU-2。将指令队列CPU-1”中的多条运行指令2-1、2-2、…、2-n，依次发送至其对应的运行设备CPU-1。将指令队列NPU-n”中的多条运行指令3-n、…、3-2、3-1，依次发送至其对应的运行设备NPU-n。将指令队列NPU-2”中的多条运行指令3’-n、…、3’-2、3’-1，依次发送至其对应的运行设备NPU-2。将队列GPU-n”中的多条运行指令4-1、4-2、…、4-n，依次发送至其对应的运行设备GPU-n。

[0236] 其中，上述运行设备CPU-2、运行设备CPU-1、运行设备NPU-n和运行设备NPU-2在接

收到指令队列之后，按照指令队列中运行指令的排列顺序，依次执行运行指令。以运行设备CPU-2为例，描述其执行所接收到的运行指令的具体过程。运行设备CPU-2包括控制模块21、执行模块22和存储模块23。其中，控制模块21包括指令存储子模块211、指令处理子模块212和存储队列子模块213，执行模块22包括依赖关系处理子模块221，详见上文关于运行设备的相关描述。

[0237] 假定根据宏指令1所生成的运行指令1-1为“@XXX……”。运行设备CPU-2在接收到运行指令1-1之后，执行运行指令1-1的过程如下：

[0238] 运行设备CPU-2的控制模块21获取数据、神经网络模型以及运行指令1-1。其中，指令存储子模块211用于存储运行指令1-1。指令处理子模块212用于对运行指令1-1进行解析，获得多个解析指令如解析指令0、解析指令1和解析指令2，并将多个解析指令发送至存储队列子模块213和执行模块22。存储队列子模块213用于存储运行指令队列，运行指令队列中包含运行设备CPU-2所需执行的解析指令0、解析指令1和解析指令2以及其他运行指令，在运行指令队列中所有指令按照执行的先后顺序依次排列。例如，获得的多个解析指令的执行的先后顺序为解析指令0、解析指令1和解析指令2，且解析指令1与解析指令0之间存在关联关系。

[0239] 运行设备CPU-2的执行模块22接收到多个解析指令后，其中的依赖关系处理子模块221判断多个解析指令之间是否存在关联关系。依赖关系处理子模块221确定出解析指令1与解析指令0存在关联关系，则将解析指令1缓存至指令存储子模块211中，并在确定解析指令0执行完毕之后，从缓存中提取出解析指令1发送至执行模块22，以供执行模块22执行。

[0240] 执行模块22接收并执行解析指令0、解析指令1和解析指令2，以完成运行指令1-1的运行。

[0241] 以上各模块的工作过程可参考上文的相关描述。

[0242] 这样，该系统可跨平台使用，适用性好，指令转换的速度快、处理效率高、出错几率低，且开发的人力、物力成本低。

[0243] 本公开提供一种机器学习运算装置，该机器学习运算装置可以包括一个或多个上述神经网络指令处理系统，用于从其他处理装置中获取待运算数据和控制信息，执行指定的机器学习运算。该机器学习运算装置可以从其他机器学习运算装置或非机器学习运算装置中获得宏指令或待执行指令，并将执行结果通过I/O接口传递给外围设备(也可称其他处理装置)。外围设备譬如摄像头，显示器，鼠标，键盘，网卡，wifi接口，服务器。当包含一个以上神经网络指令处理系统时，神经网络指令处理系统间可以通过特定的结构进行链接并传输数据，譬如，通过PCIE总线进行互联并传输数据，以支持更大规模的神经网络的运算。此时，可以共享同一控制系统，也可以有各自独立的控制系统；可以共享内存，也可以每个加速器有各自的内存。此外，其互联方式可以是任意互联拓扑。

[0244] 该机器学习运算装置具有较高的兼容性，可通过PCIE接口与各种类型的服务器相连接。

[0245] 图4a示出根据本公开一实施例的组合处理装置的框图。如图4a所示，该组合处理装置包括上述机器学习运算装置、通用互联接口和其他处理装置。机器学习运算装置与其他处理装置进行交互，共同完成用户指定的操作。

[0246] 其他处理装置，包括中央处理器CPU、图形处理器GPU、神经网络处理器等通用/专

用处理器中的一种或以上的处理器类型。其他处理装置所包括的处理器数量不做限制。其他处理装置作为机器学习运算装置与外部数据和控制的接口，包括数据搬运，完成对本机器学习运算装置的开启、停止等基本控制；其他处理装置也可以和机器学习运算装置协作共同完成运算任务。

[0247] 通用互联接口，用于在机器学习运算装置与其他处理装置间传输数据和控制指令。该机器学习运算装置从其他处理装置中获取所需的输入数据，写入机器学习运算装置片上的存储装置；可以从其他处理装置中获取控制指令，写入机器学习运算装置片上的控制缓存；也可以读取机器学习运算装置的存储模块中的数据并传输给其他处理装置。

[0248] 图4b示出根据本公开一实施例的组合处理装置的框图。在一种可能的实现方式中，如图4b所示，该组合处理装置还可以包括存储装置，存储装置分别与机器学习运算装置和所述其他处理装置连接。存储装置用于保存在机器学习运算装置和所述其他处理装置的数据，尤其适用于所需要运算的数据在本机器学习运算装置或其他处理装置的内部存储中无法全部保存的数据。

[0249] 该组合处理装置可以作为手机、机器人、无人机、视频监控设备等设备的SOC片上系统，有效降低控制部分的核心面积，提高处理速度，降低整体功耗。此情况时，该组合处理装置的通用互联接口与设备的某些部件相连接。某些部件譬如摄像头，显示器，鼠标，键盘，网卡，wifi接口。

[0250] 本公开提供一种机器学习芯片，该芯片包括上述机器学习运算装置或组合处理装置。

[0251] 本公开提供一种机器学习芯片封装结构，该机器学习芯片封装结构包括上述机器学习芯片。

[0252] 本公开提供一种板卡，图5示出根据本公开一实施例的板卡的结构示意图。如图5所示，该板卡包括上述机器学习芯片封装结构或者上述机器学习芯片。板卡除了包括机器学习芯片389以外，还可以包括其他的配套部件，该配套部件包括但不限于：存储器件390、接口装置391和控制器件392。

[0253] 存储器件390与机器学习芯片389(或者机器学习芯片封装结构内的机器学习芯片)通过总线连接，用于存储数据。存储器件390可以包括多组存储单元393。每一组存储单元393与机器学习芯片389通过总线连接。可以理解，每一组存储单元393可以是DDR SDRAM(英文：Double Data Rate SDRAM，双倍速率同步动态随机存储器)。

[0254] DDR不需要提高时钟频率就能加倍提高SDRAM的速度。DDR允许在时钟脉冲的上升沿和下降沿读出数据。DDR的速度是标准SDRAM的两倍。

[0255] 在一个实施例中，存储器件390可以包括4组存储单元393。每一组存储单元393可以包括多个DDR4颗粒(芯片)。在一个实施例中，机器学习芯片389内部可以包括4个72位DDR4控制器，上述72位DDR4控制器中64bit用于传输数据，8bit用于ECC校验。可以理解，当每一组存储单元393中采用DDR4-3200颗粒时，数据传输的理论带宽可达到25600MB/s。

[0256] 在一个实施例中，每一组存储单元393包括多个并联设置的双倍速率同步动态随机存储器。DDR在一个时钟周期内可以传输两次数据。在机器学习芯片389中设置控制DDR的控制器，用于对每个存储单元393的数据传输与数据存储的控制。

[0257] 接口装置391与机器学习芯片389(或者机器学习芯片封装结构内的机器学习芯

片)电连接。接口装置391用于实现机器学习芯片389与外部设备(例如服务器或计算机)之间的数据传输。例如在一个实施例中,接口装置391可以为标准PCIE接口。比如,待处理的数据由服务器通过标准PCIE接口传递至机器学习芯片289,实现数据转移。优选的,当采用PCIE 3.0X 16接口传输时,理论带宽可达到16000MB/s。在另一个实施例中,接口装置391还可以是其他的接口,本公开并不限制上述其他的接口的具体表现形式,接口装置能够实现转接功能即可。另外,机器学习芯片的计算结果仍由接口装置传送回外部设备(例如服务器)。

[0258] 控制器件392与机器学习芯片389电连接。控制器件392用于对机器学习芯片389的状态进行监控。具体的,机器学习芯片389与控制器件392可以通过SPI接口电连接。控制器件392可以包括单片机(Micro Controller Unit,MCU)。如机器学习芯片389可以包括多个处理芯片、多个处理核或多个处理电路,可以带动多个负载。因此,机器学习芯片389可以处于多负载和轻负载等不同的工作状态。通过控制器件可以实现对机器学习芯片中多个处理芯片、多个处理和/或多个处理电路的工作状态的调控。

[0259] 本公开提供一种电子设备,该电子设备包括上述机器学习芯片或板卡。

[0260] 电子设备可以包括数据处理装置、机器人、电脑、打印机、扫描仪、平板电脑、智能终端、手机、行车记录仪、导航仪、传感器、摄像头、服务器、云端服务器、相机、摄像机、投影仪、手表、耳机、移动存储、可穿戴设备、交通工具、家用电器、和/或医疗设备。

[0261] 交通工具可以包括飞机、轮船和/或车辆。家用电器可以包括电视、空调、微波炉、冰箱、电饭煲、加湿器、洗衣机、电灯、燃气灶、油烟机。医疗设备可以包括核磁共振仪、B超仪和/或心电图仪。

[0262] 图6示出根据本公开一实施例的神经网络指令处理方法的流程图。如图6所示,该方法应用于上述包括指令生成设备和运行设备的神经网络指令处理系统,该方法包括步骤S41和步骤S42。

[0263] 在步骤S41中,通过指令生成设备根据接收到的宏指令,确定执行宏指令的运行设备,并根据宏指令和运行设备,生成运行指令。

[0264] 在步骤S42中,通过运行设备获取数据、神经网络模型以及运行指令,对运行指令进行解析,获得多个解析指令,并根据数据执行多个解析指令,得到执行结果。

[0265] 在一种可能的实现方式中,步骤S41可以包括:在确定宏指令中包含指定设备的标识,且指定设备的资源满足执行宏指令的执行条件时,将指定设备确定为运行设备。其中,执行条件可以包括:指定设备中包含与宏指令相对应的指令集。

[0266] 在一种可能的实现方式中,该方法还可以包括:获取备选设备的资源信息。

[0267] 其中,步骤S41还可以包括:在确定宏指令中不包含指定设备的标识时,根据接收到的宏指令和备选设备的资源信息,从备选设备中确定出用于执行宏指令的运行设备。其中,资源信息可以包括备选设备所包含的指令集。

[0268] 在一种可能的实现方式中,步骤S41还可以包括:在确定宏指令中包含指定设备的标识,且指定设备的资源不满足执行宏指令的执行条件时,根据宏指令和备选设备的资源信息,确定运行设备。

[0269] 在一种可能的实现方式中,宏指令可以包含输入量和输出量中的至少一项,步骤S41中根据宏指令和运行设备,生成运行指令,可以包括:确定宏指令的数据量,根据宏指令

的数据量、宏指令和运行设备的资源信息，生成运行指令。其中，数据量是根据输入量和输出量中的至少一项确定的，运行设备的资源信息还包括存储容量、剩余存储容量的至少一项。

[0270] 在一种可能的实现方式中，根据宏指令的数据量、宏指令和运行设备的资源信息，生成运行指令，可以包括：在确定运行设备为一个，且运行设备的资源不满足执行宏指令的容量条件时，根据运行设备的运行数据量和数据量将宏指令拆分成多条运行指令，以使运行设备依次执行多条运行指令。其中，运行设备的运行数据量可以是根据运行设备的资源信息确定的，每条运行指令可以包含运行输入量和运行输出量中的至少一项，运行输入量和运行输出量是根据运行数据量确定的。

[0271] 在一种可能的实现方式中，根据宏指令的数据量、宏指令和运行设备的资源信息，生成运行指令，可以包括：在确定运行设备为多个时，根据每个运行设备的运行数据量和数据量对宏指令进行拆分，生成对应于每个运行设备的运行指令。其中，每个运行设备的运行数据量可以是根据每个运行设备的资源信息确定的，运行指令可以包含运行输入量和运行输出量中的至少一项，运行输入量和运行输出量是根据执行运行指令的运行设备的运行数据量确定的。

[0272] 在一种可能的实现方式中，该方法还可以包括：通过指令生成设备根据队列排序规则对运行指令进行排序，根据排序后的运行指令构建与运行设备相对应的指令队列。

[0273] 在一种可能的实现方式中，该方法还可以包括：通过指令生成设备接收待执行指令，根据确定的指定设备的标识和待执行指令生成宏指令。

[0274] 在一种可能的实现方式中，该方法还可以包括：通过指令生成设备将运行指令发送至运行设备，以使运行设备执行运行指令。

[0275] 在一种可能的实现方式中，通过指令生成设备将运行指令发送至运行设备，以使运行设备执行运行指令，可以包括：

[0276] 根据运行指令生成汇编文件；

[0277] 将汇编文件翻译成二进制文件；

[0278] 将二进制文件发送至运行设备，以使运行设备根据二进制文件执行运行指令。

[0279] 在一种可能的实现方式中，该方法还可以包括：通过运行设备存储数据以及数据中的标量数据。其中，运行设备包括存储模块，存储模块包括寄存器、缓存中任意组合，缓存包括高速暂存缓存，缓存，用于存储数据；寄存器，用于存储数据中标量数据。

[0280] 在一种可能的实现方式中，该方法还可以包括：

[0281] 通过运行设备存储运行指令；

[0282] 通过运行设备对运行指令进行解析，得到多个解析指令；

[0283] 通过运行设备存储运行指令队列，运行指令队列包括运行指令和多个解析指令，运行指令队列运行指令和多个解析指令按照被执行的先后顺序依次排列。

[0284] 在一种可能的实现方式中，该方法还可以包括：

[0285] 通过运行设备在确定第一解析指令与第一解析指令之前的第零解析指令存在关联关系时，缓存第一解析指令，在第零解析指令执行完毕后，执行缓存的第一解析指令。

[0286] 其中，第一解析指令与第一解析指令之前的第零解析指令存在关联关系包括：存储第一解析指令所需数据的第一存储地址区间与存储第零解析指令所需数据的第零存储

地址区间具有重叠的区域。

[0287] 在一种可能的实现方式中，运行设备可以为CPU、GPU和NPU中的其中一种或任意组合。

[0288] 在一种可能的实现方式中，指令生成设备设置于CPU和/或NPU中。

[0289] 在一种可能的实现方式中，宏指令可以包括以下指令中的至少一种：计算宏指令、控制宏指令和数据搬运宏指令。

[0290] 其中，计算宏指令可以包括神经网络计算宏指令、向量逻辑计算宏指令、矩阵向量计算宏指令、标量计算宏指令和标量逻辑计算宏指令中的至少一种。控制宏指令可以包括无条件跳转宏指令和有条件跳转宏指令中的至少一种。数据搬运宏指令可以包括读宏指令和写宏指令中的至少一种。读宏指令可以包括读神经元宏指令、读突触宏指令和读标量宏指令中的至少一种。写宏指令可以包括写神经元宏指令、写突触宏指令和写标量宏指令中的至少一种。

[0291] 在一种可能的实现方式中，宏指令可以包含以下选项中的至少一项：用于执行宏指令的指定设备的标识、操作类型、输入地址、输出地址、输入量、输出量、操作数和指令参数。运行指令可以包含以下选项中的至少一项：操作类型、输入地址、输出地址、操作数和指令参数。

[0292] 本公开实施例所提供的神经网络指令处理方法，通过指令生成设备根据接收到的宏指令，确定执行宏指令的运行设备，并根据宏指令和运行设备，生成运行指令；通过运行设备获取数据、神经网络模型以及运行指令，对运行指令进行解析，获得多个解析指令，并根据数据执行多个解析指令，得到执行结果。该方法可跨平台使用，适用性好，指令转换的速度快、处理效率高、出错几率低，且开发的人力、物力成本低。

[0293] 需要说明的是，对于前述的各方法实施例，为了简单描述，故将其都表述为一系列的动作组合，但是本领域技术人员应该知悉，本申请并不受所描述的动作顺序的限制，因为依据本申请，某些步骤可以采用其他顺序或者同时进行。其次，本领域技术人员也应该知悉，说明书中所描述的实施例均属于可选实施例，所涉及的动作和模块并不一定是本公开所必须的。

[0294] 在上述实施例中，对各个实施例的描述都各有侧重，某个实施例中没有详述的部分，可以参见其他实施例的相关描述。

[0295] 在本公开所提供的实施例中，应该理解到，所揭露的系统、装置，可通过其它的方式实现。例如，以上所描述的系统、装置实施例仅仅是示意性的，例如设备、装置、模块的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个模块可以结合或者可以集成到另一个系统或装置，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，设备、装置或模块的间接耦合或通信连接，可以是电性或其它的形式。

[0296] 作为分离部件说明的模块可以是或者也可以不是物理上分开的，作为模块显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。

[0297] 另外，在本公开各个实施例中的各功能模块可以集成在一个处理单元中，也可以是各个模块单独物理存在，也可以两个或两个以上模块集成在一个模块中。上述集成的模

块既可以采用硬件的形式实现，也可以采用软件程序模块的形式实现。

[0298] 集成的模块如果以软件程序模块的形式实现并作为独立的产品销售或使用时，可以存储在一个计算机可读取存储器中。基于这样的理解，本公开的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的全部或部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储器中，包括若干指令用以使得一台计算机设备(可为个人计算机、服务器或者网络设备等)执行本公开各个实施例所述方法的全部或部分步骤。而前述的存储器包括：U盘、只读存储器(ROM，Read-Only Memory)、随机存取存储器(RAM，Random Access Memory)、移动硬盘、磁碟或者光盘等各种可以存储程序代码的介质。

[0299] 本领域普通技术人员可以理解上述实施例的各种方法中的全部或部分步骤是可以通过程序来指令相关的硬件来完成，该程序可以存储于一计算机可读存储器中，存储器可以包括：闪存盘、只读存储器(英文：Read-Only Memory，简称：ROM)、随机存取器(英文：Random Access Memory，简称：RAM)、磁盘或光盘等。

[0300] 以上对本申请实施例进行了详细介绍，本文中应用了具体个例对本申请的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本申请的方法及其核心思想；同时，对于本领域的一般技术人员，依据本申请的思想，在具体实施方式及应用范围上均会有改变之处，综上所述，本说明书内容不应理解为对本申请的限制。

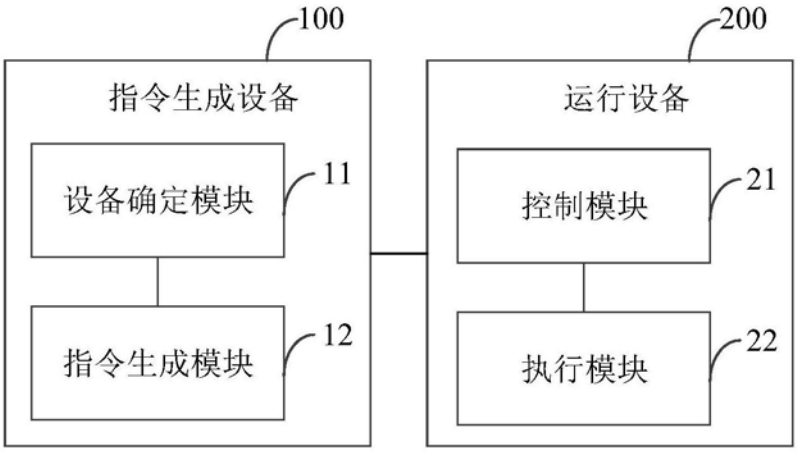


图1

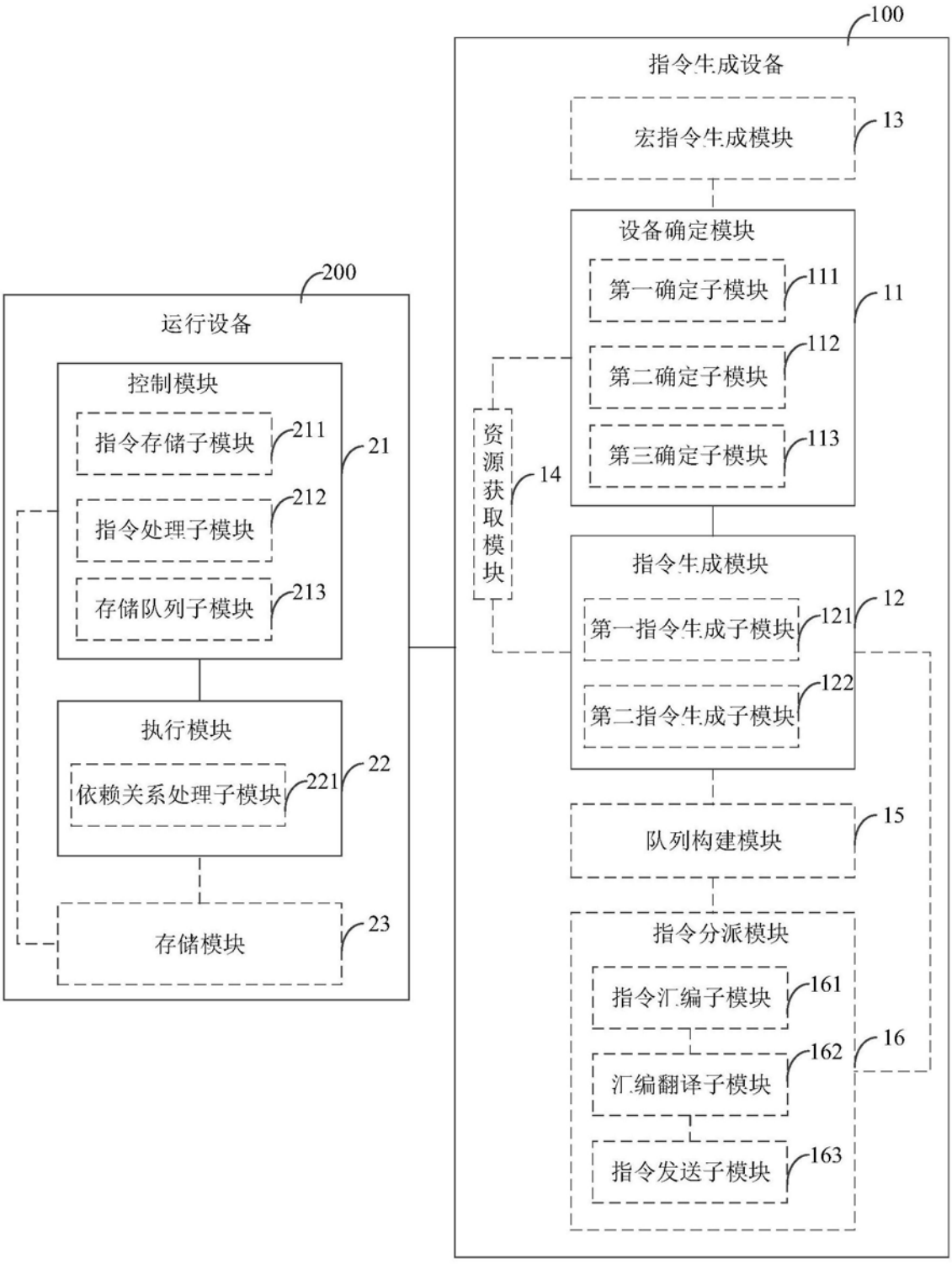

100
指令生成设备
宏指令生成模块
13
设备确定模块
第一确定子模块
111
11
第二确定子模块
112
第三确定子模块
113
资源获取模块
14
指令生成模块
第一指令生成子模块
121
12
第二指令生成子模块
122
队列构建模块
15
指令分派模块
指令汇编子模块
161
汇编翻译子模块
162
16
指令发送子模块
163
200
运行设备
控制模块
指令存储子模块
211
21
指令处理子模块
212
存储队列子模块
213
执行模块
依赖关系处理子模块
221
22
存储模块
23

图2

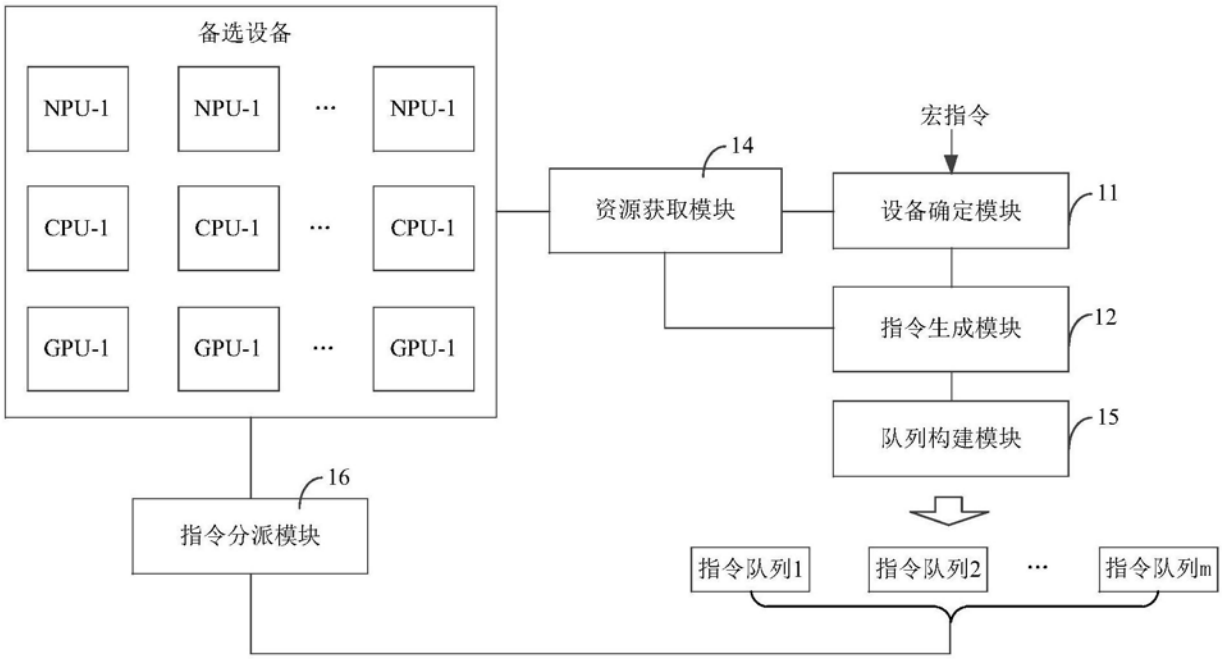

备选设备
NPU-1
NPU-1
…
NPU-1
CPU-1
CPU-1
…
CPU-1
GPU-1
GPU-1
…
GPU-1
资源获取模块
14
宏指令
设备确定模块
11
指令生成模块
12
队列构建模块
15
指令分派模块
16
指令队列1
指令队列2
…
指令队列m

图3a

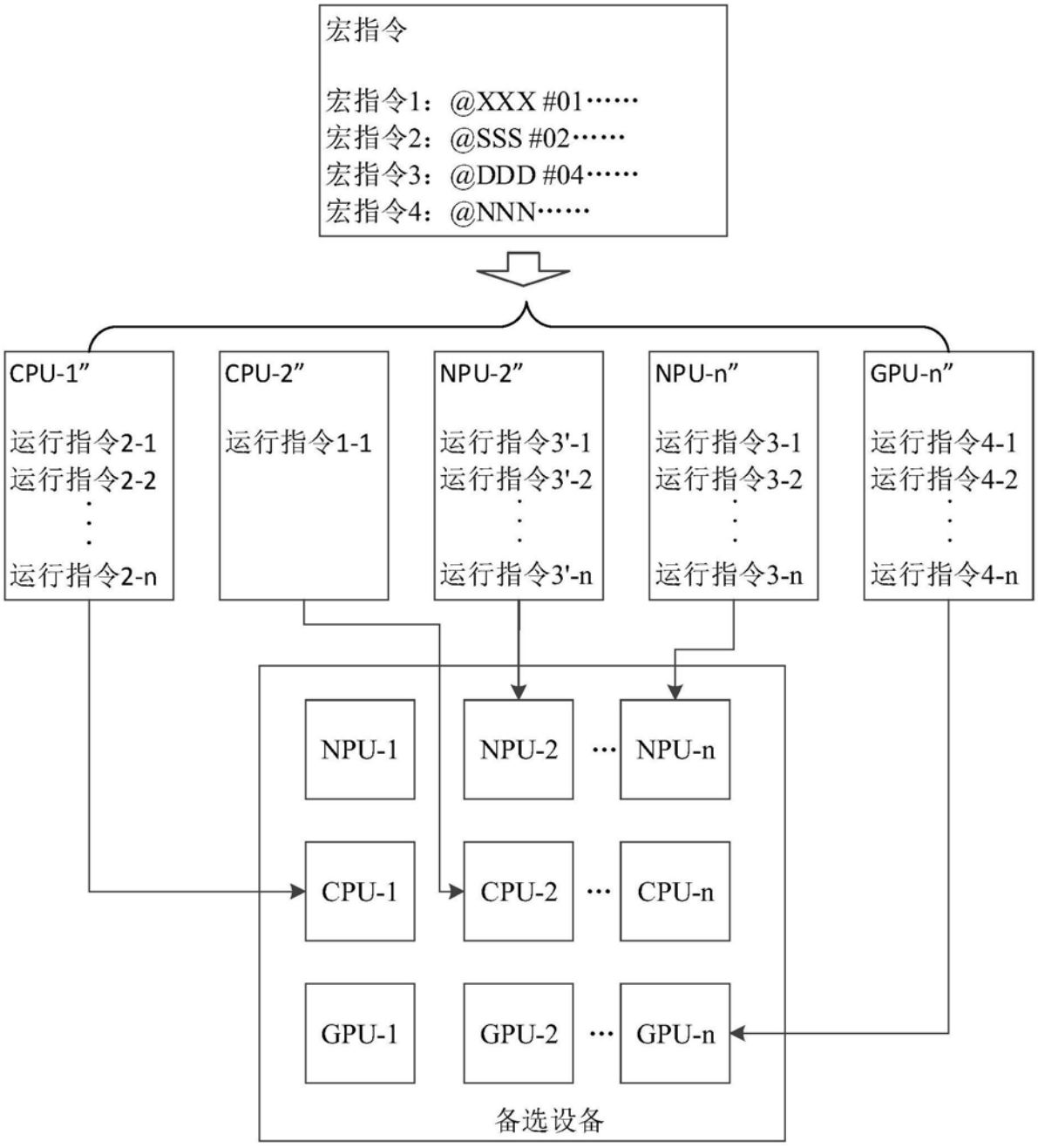

宏指令
宏指令1：@XXX #01……
宏指令2：@SSS #02……
宏指令3：@DDD #04……
宏指令4：@NNN……
CPU-1”
运行指令2-1
运行指令2-2
运行指令2-n
CPU-2”
运行指令1-1
NPU-2”
运行指令3'-1
运行指令3'-2
运行指令3'-n
NPU-n”
运行指令3-1
运行指令3-2
运行指令3-n
GPU-n”
运行指令4-1
运行指令4-2
运行指令4-n
NPU-1
NPU-2
NPU-n
CPU-1
CPU-2
CPU-n
GPU-1
GPU-2
GPU-n
备选设备

图3b

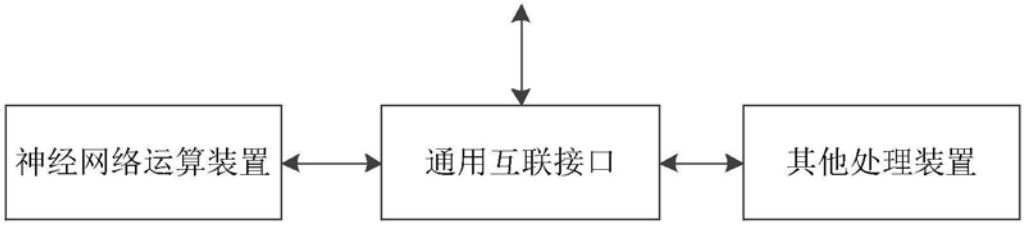

神经网络运算装置
通用互联接口
其他处理装置

图4a

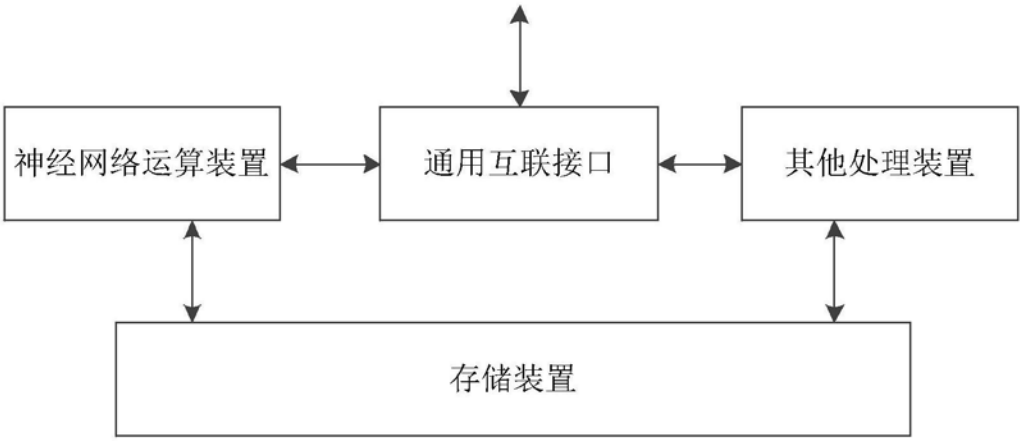


图4b

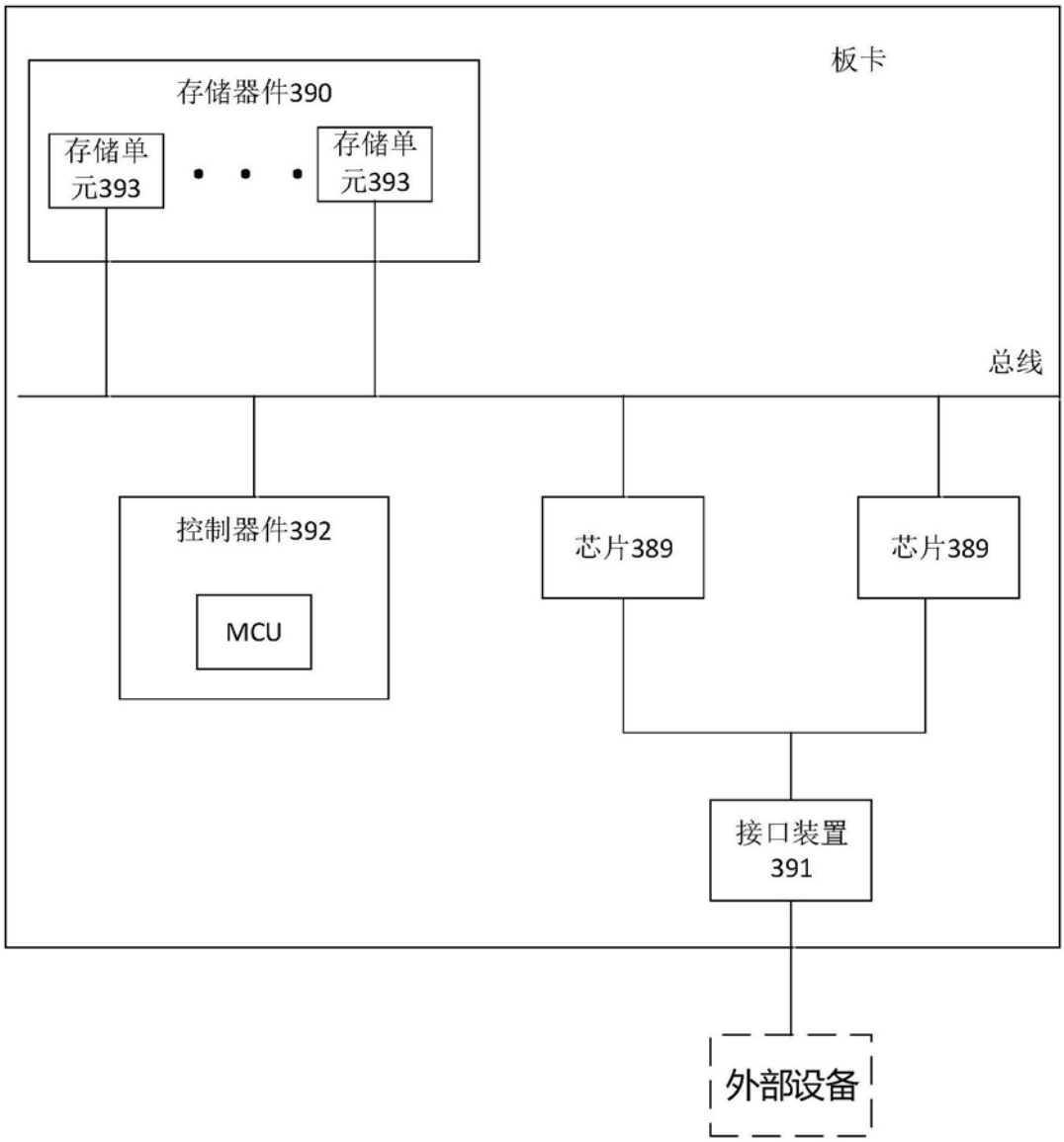

板卡
存储器件390
存储单元393
存储单元393
总线
控制器件392
MCU
芯片389
芯片389
接口装置391
外部设备

图5


通过指令生成设备根据接收到的宏指令，确定执行宏指令的运行设备，并根据宏指令和运行设备，生成运行指令
S41
通过运行设备获取数据、神经网络模型以及运行指令，对运行指令进行解析，获得多个解析指令，并根据数据执行多个解析指令，得到执行结果
S42

图6