(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 118821966 A
(43)申请公布日 2024.10.22

(21)申请号 202410462673.2
(22)申请日 2024.04.17
(30)优先权数据
18/301,816 2023.04.17 US
(71)申请人 元平台公司
地址 美国加利福尼亚州
(72)发明人 考沙尔·甘地 奥利维亚·吴
索希尔·加拉希
托马斯·马克·乌尔里奇
卡西姆·S·杜德库拉
埃达·萨欣
阿卜杜勒卡迪尔·乌特库·迪里尔
(74)专利代理机构 北京中博世达专利商标代理有限公司 11274
专利代理师 于宁娜 程艳茹
(51)Int.Cl.
G06N 20/00(2019.01)

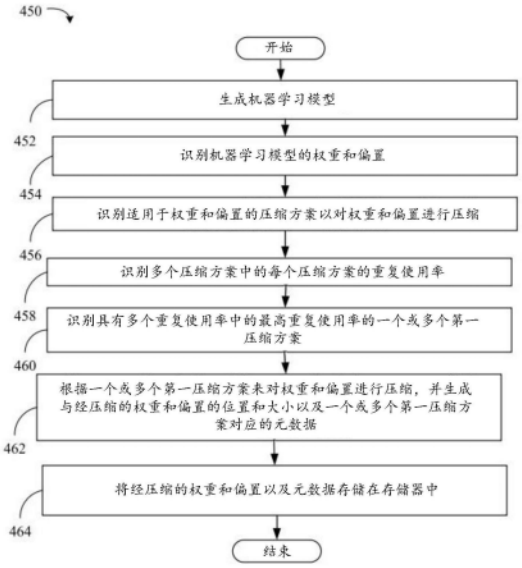
权利要求书4页 说明书32页 附图17页

(54)发明名称

对与机器学习模型相关联的数据的选择性和灵活性压缩

(57)摘要

本公开提供了对与机器学习模型相关联的数据的选择性和灵活性压缩。系统、装置和方法提供了技术,该技术基于第一压缩方案对第一数据进行压缩以生成第二数据,其中,第一数据与第一机器学习模型相关联。该技术将第二数据存储到存储器中;根据第一数据是基于第一压缩方案而被压缩来调整查找表的第一条目以对应于第一压缩方案;在第一机器学习模型的执行期间,向处理阵列的处理元件提供来自存储器的第二数据;以及在处理阵列处,基于查找表对第二数据进行解压缩以获得第一数据。



CN 118821966 A

1.至少一个计算机可读存储介质,所述至少一个计算机可读存储介质包括指令集,所述指令集在被计算设备执行时使所述计算设备:

基于第一压缩方案对第一数据进行压缩以生成第二数据,其中,所述第一数据与第一机器学习模型相关联;

将所述第二数据存储到存储器中;

根据所述第一数据是基于所述第一压缩方案而被压缩,来调整查找表的第一条目以对应于所述第一压缩方案;

在所述第一机器学习模型的执行期间,向处理阵列的处理元件提供来自所述存储器的所述第二数据;以及

在所述处理阵列处,基于所述查找表对所述第二数据进行解压缩以获得所述第一数据。

2.根据权利要求1所述的至少一个计算机可读存储介质,其中,所述指令集在被执行时使所述计算设备:

基于第二压缩方案对第三数据进行压缩以生成第四数据,其中,所述第三数据与第二机器学习模型相关联;

将所述第四数据存储到所述存储器中;

根据所述第三数据是基于所述第二压缩方案而被压缩,来调整所述查找表的第二条目以对应于所述第二压缩方案;

在所述第二机器学习模型的执行期间,向所述处理阵列提供来自所述存储器的所述第三数据;以及

在所述处理阵列处,基于所述查找表对所述第四数据进行解压缩以获得所述第三数据。

3.根据权利要求2所述的至少一个计算机可读存储介质,其中,所述指令集在被执行时使所述计算设备:

用所述第二条目替换所述查找表中的所述第一条目。

4.根据权利要求2所述的至少一个计算机可读存储介质,其中,所述指令集在被执行时使所述计算设备:

将所述第二条目插入到所述查找表中,使得所述查找表包括所述第一条目和所述第二条目。

5.根据权利要求1所述的至少一个计算机可读存储介质,其中,所述指令集在被执行时使所述计算设备:

基于所述第一压缩方案,识别所述第一数据的第一部分能够压缩一量;以及

基于所述量满足压缩阈值,确定要用所述第一压缩方案对所述第一部分进行压缩。

6.根据权利要求1所述的至少一个计算机可读存储介质,其中,所述指令集在被执行时使所述计算设备:

确定用所述第一压缩方案对所述第一数据进行压缩将所述第一数据的大小减小到小于或等于所述存储器的可用存储器空间;

预测所述第一压缩方案要被应用至少第一次数;

确定所述至少第一次数满足重复使用阈值;

基于所述至少第一次数满足所述重复使用阈值,和用所述第一压缩方案对所述第一数据进行压缩将所述第一数据的大小减小到小于或等于所述可用存储器空间,来选择所述第一压缩方案;

生成元数据以将所述第一压缩方案与所述第二数据相关联;以及

将所述第二数据与所述元数据相关联地存储到所述存储器中。

7.根据权利要求1所述的至少一个计算机可读存储介质,其中:

所述第一数据包括权重或偏置中的一者或多者;

所述第一压缩方案是非对称数字系统压缩方案;

所述第一数据是在所述第一机器模型被执行之前基于所述第一压缩方案来压缩的;以及

为了对所述第二数据进行解压缩,所述处理阵列用于使用所述处理元件对所述第二数据进行解压缩。

8.一种系统,包括:

一个或多个处理器;以及

第一存储器,所述第一存储器耦接到所述一个或多个处理器,所述第一存储器包括能够由所述一个或多个处理器执行的指令,所述一个或多个处理器在执行所述指令时能够操作以:

基于第一压缩方案对第一数据进行压缩以生成第二数据,其中,所述第一数据与第一机器学习模型相关联;

将所述第二数据存储到第二存储器中;

根据所述第一数据是基于所述第一压缩方案而被压缩,来调整查找表的第一条目以对应于所述第一压缩方案;

在所述第一机器学习模型的执行期间,向处理阵列的处理元件提供来自所述第二存储器的所述第二数据;以及

在所述处理阵列处,基于所述查找表对所述第二数据进行解压缩以获得所述第一数据。

9.根据权利要求8所述的系统,其中,所述一个或多个处理器在执行所述指令时,还能够操作以:

基于第二压缩方案对第三数据进行压缩以生成第四数据,其中,所述第三数据与第二机器学习模型相关联;

将所述第四数据存储到所述第二存储器中;

根据所述第三数据是基于所述第二压缩方案而被压缩,来调整所述查找表的第二条目以对应于所述第二压缩方案;

在所述第二机器学习模型的执行期间,向所述处理阵列提供来自所述第二存储器的所述第三数据;以及

在所述处理阵列处,基于所述查找表对所述第四数据进行解压缩以获得所述第三数据。

10.根据权利要求9所述的系统,其中,所述一个或多个处理器在执行所述指令时还能够操作以:

用所述第二条目替换所述查找表中的所述第一条目。

11.根据权利要求9所述的系统,其中,所述一个或多个处理器在执行所述指令时还能够操作以:

将所述第二条目插入到所述查找表中,使得所述查找表包括所述第一条目和所述第二条目。

12.根据权利要求8所述的系统,其中,所述一个或多个处理器在执行所述指令时还能够操作以:

基于所述第一压缩方案,识别所述第一数据的第一部分能够压缩一量;以及

基于所述量满足压缩阈值,确定要用所述第一压缩方案对所述第一部分进行压缩。

13.根据权利要求8所述的系统,其中,所述一个或多个处理器在执行所述指令时还能够操作以:

确定用所述第一压缩方案对所述第一数据进行压缩将所述第一数据的大小减小到小于或等于所述第二存储器的可用存储器空间;

预测所述第一压缩方案要被应用至少第一次数;

确定所述至少第一次数满足重复使用阈值;

基于所述至少第一次数满足所述重复使用阈值,和用所述第一压缩方案对所述第一数据进行压缩将所述第一数据的大小减小到小于或等于所述可用存储器空间,来选择所述第一压缩方案;

生成元数据以将所述第一压缩方案与所述第二数据相关联;以及

将所述第二数据与所述元数据相关联地存储到所述第二存储器中。

14.根据权利要求8所述的系统,其中,

所述第一数据包括权重或偏置中的一者或多者;

所述第一压缩方案是非对称数字系统压缩方案;

所述第一数据是在所述第一机器模型被执行之前基于所述第一压缩方案来压缩的;以及

为了对所述第二数据进行解压缩,所述处理阵列用于使用所述处理元件对所述第二数据进行解压缩。

15.一种方法,包括:

基于第一压缩方案对第一数据进行压缩以生成第二数据,其中,所述第一数据与第一机器学习模型相关联;

将所述第二数据存储到存储器中;

根据所述第一数据是基于所述第一压缩方案而被压缩,来调整查找表的第一条目以对应于所述第一压缩方案;

在所述第一机器学习模型的执行期间,向处理阵列的处理元件提供来自所述存储器的所述第二数据;以及

在所述处理阵列处,基于所述查找表对所述第二数据进行解压缩以获得所述第一数据。

16.根据权利要求15所述的方法,还包括:

基于第二压缩方案对第三数据进行压缩以生成第四数据,其中,所述第三数据与第二

机器学习模型相关联;

将所述第四数据存储到所述存储器中;

根据所述第三数据是基于所述第二压缩方案而被压缩,来调整所述查找表的第二条目以对应于所述第二压缩方案;

在执行所述第二机器学习模型期间,向所述处理阵列提供来自所述存储器的所述第三数据;以及

在所述处理阵列处,基于所述查找表对所述第四数据进行解压缩以获得所述第三数据。

17.根据权利要求16所述的方法,还包括:

用所述第二条目替换所述查找表中的所述第一条目。

18.根据权利要求16所述的方法,还包括:

将所述第二条目插入到所述查找表中,使得所述查找表包括所述第一条目和所述第二条目。

19.根据权利要求15所述的方法,还包括:

基于所述第一压缩方案,识别所述第一数据的第一部分能够压缩一量;以及

基于所述量满足压缩阈值,确定要用所述第一压缩方案对所述第一部分进行压缩。

20.根据权利要求15所述的方法,还包括:

确定用所述第一压缩方案对所述第一数据进行压缩将所述第一数据的大小减小到小于或等于所述存储器的可用存储器空间;

预测所述第一压缩方案要被应用至少第一次数;

确定所述至少第一次数满足重复使用阈值;

基于所述至少第一次数满足所述重复使用阈值,和用所述第一压缩方案对所述第一数据进行压缩将所述第一数据的大小减小到小于或等于所述可用存储器空间,来选择所述第一压缩方案;

生成元数据以将所述第一压缩方案与所述第二数据相关联;以及

将所述第二数据与所述元数据相关联地存储到所述存储器中,

其中:

所述第一数据包括权重或偏置中的一者或多者;

所述第一压缩方案是非对称数字系统压缩方案;

所述第一数据是在所述第一机器模型被执行之前基于所述第一压缩方案来压缩的;以及

为了对所述第二数据进行解压缩,所述处理阵列用于使用所述处理元件对所述第二数据进行解压缩。

对与机器学习模型相关联的数据的选择性和灵活性压缩

[0001] 相关申请的交叉引用

[0002] 本申请要求于2023年4月17日提交的申请号为18/301,816的美国非临时申请的优先权,该申请通过引用并入本文。

技术领域

[0003] 本公开总体上涉及对与机器学习模型相关联的数据的选择性和灵活性压缩。

背景技术

[0004] 机器学习模型可以基于训练数据来训练。机器学习模型可以执行迭代训练过程以基于训练数据(例如,经由梯度)来训练机器学习模型。机器学习模型在经过训练后,可以执行人工智能(artificial intelligence,AI)推理。AI推理可以是使用经训练的机器学习模型基于输入数据做出预测或决策的过程。例如,推理可以是使用经训练的模型进行预测的过程。

发明内容

[0005] 一个或多个计算机的系统可以被配置为通过将软件、固件、硬件或它们的组合安装在该系统上来执行特定操作或动作,所述软件、固件、硬件或它们的组合运行时使该系统执行动作。一个或多个计算机程序可以被配置为通过包括指令来执行特定操作或动作,所述指令在被数据处理装置执行时使该数据处理装置执行这些动作。

[0006] 一个总体方面包括至少一个计算机可读存储介质,该至少一个计算机可读存储介质包括指令集,该指令集在被计算设备执行时使得该计算设备基于第一压缩方案对第一数据进行压缩以生成第二数据,其中,该第一数据与第一机器学习模型相关联。这些指令在被执行时使得该计算设备将第二数据存储到存储器中,并根据第一数据是基于第一压缩方案而被压缩,来调整查找表的第一条目以对应于第一压缩方案。这些指令在被执行时使得计算设备在第一机器学习模型的执行期间向处理阵列的处理元件提供来自该存储器的第二数据,以及在该处理阵列处基于查找表对第二数据进行解压缩以获得第一数据。该方面的其它示例包括对应的计算机系统、装置和在一个或多个计算机存储设备上记录的计算机程序,每个计算机系统、装置和计算机程序被配置为执行这些方法的动作。

[0007] 一个总体方面包括一种系统,该系统包括一个或多个处理器以及耦接到该一个或多个处理器的存储器,该存储器包括能够由该一个或多个处理器执行的指令,该一个或多个处理器在执行指令时能够操作以:基于第一压缩方案对第一数据进行压缩以生成第二数据,其中,第一数据与第一机器学习模型相关联。该一个或多个处理器在执行指令时还能够操作以:将第二数据存储到第二存储器中,根据第一数据是基于第一压缩方案而被压缩,来调整查找表的第一条目以对应于第一压缩方案,以及在第一机器学习模型的执行期间,向处理阵列的处理元件提供来自第二存储器的第二数据。该一个或多个处理器在执行指令时还能够操作以:在处理阵列处,基于查找表对第二数据进行解压缩以获得第一数据。该方面

的其它示例包括对应的计算机系统、装置和在一个或多个计算机存储设备上记录的计算机程序,每个计算机系统、装置和计算机程序被配置为执行这些方法的动作。

[0008] 一个总体方面包括一种方法,该方法包括基于第一压缩方案对第一数据进行压缩以生成第二数据,其中,第一数据与第一机器学习模型相关联。该方法还包括将第二数据存储到存储器中。该方法还包括根据第一数据是基于第一压缩方案而被压缩,来调整查找表的第一条目以对应于第一压缩方案。该方法还包括在第一机器学习模型的执行期间,向处理阵列的处理元件提供来自存储器的第二数据。该方法还包括在处理阵列处,基于查找表对第二数据进行解压缩以获得第一数据。该方面的其它示例包括对应的计算机系统、装置和在一个或多个计算机存储设备上记录的计算机程序,每个计算机系统、装置和计算机程序被配置为执行这些方法的动作。

附图说明

[0009] 对于本领域技术人员而言,通过阅读以下说明书和所附权利要求书并通过参考以下附图,示例的各种优点将变得显而易见,在这些附图中:

[0010] 图1例示了根据本公开示例的执行对机器学习模型数据的压缩和解压缩的计算架构;

[0011] 图2是根据本公开示例的用于对与机器学习模型相关联的数据进行压缩并进行存储的方法的示例的流程图;

[0012] 图3是根据本公开示例的用于对机器学习数据进行压缩和解压缩的方法的示例的流程图;

[0013] 图4是根据本公开示例的用于对机器学习数据进行压缩和解压缩的方法的示例的流程图;

[0014] 图5是根据本公开示例的用于对机器学习模型的数据进行压缩的方法的示例的流程图;

[0015] 图6A例示了根据本公开示例的经解压缩的缓冲区;

[0016] 图6B例示了根据本公开示例的存储器布局;

[0017] 图7例示了根据本公开示例的每个块(chunk)在被压缩之前所经历的过程;

[0018] 图8例示了示出了根据本公开示例的对指令或元数据的分解的示意图;

[0019] 图9A例示了根据本公开示例的以示例列主遍历模式来遍历的经压缩的缓冲区的布局的示意图;

[0020] 图9B例示了根据本公开示例在没有启用交换比特的情况下将元数据存储在存储器中的布局;

[0021] 图9C例示了根据本公开示例的在启用了交换比特的情况下在存储器中的元数据的布局;

[0022] 图10例示了根据本公开示例的解压缩架构;

[0023] 图11例示了根据本公开示例的高级解压缩流水线;

[0024] 图12例示了根据本公开示例的与社交网络系统相关联的示例网络环境;

[0025] 图13例示了根据本公开示例的示例社交图谱;以及

[0026] 图14例示了根据本公开示例的示例计算机系统。

具体实施方式

[0027] 机器学习模型可能会产生巨大的开销和性能成本。例如,机器学习模型可以包括权重和偏置(bias)。在推理期间,机器学习模型可以接收数据,这些数据由机器学习模型处理以形成预测、分类、决策等。数据可以实时地流式传输到机器学习模型中。与机器学习模型相关联的数据(例如,权重、偏置、输入数据等)可能具有使存储和检索复杂化的显著大小。例如,芯片可以利用机器学习模型执行推理过程。芯片的片上存储器的可用存储器空间可能不能存储所有的数据,从而导致从存储位置(例如,诸如存储器、硬盘驱动器或数据库等片外存储装置)获取数据以及从存储器淘汰(eviction)数据的时间较长。这样做会增加延迟(例如,等待数据的延迟)和计算资源(例如,用于淘汰数据和检索数据的资源)。例如,将此类数据传输到存储器可能会消耗大量带宽。在一些示例中,该机器学习模型可能在多个芯片上执行。这样做会增加计算资源消耗,因为利用了多个芯片并且还执行了用于将该多个芯片的输出进行组合的操作。

[0028] 本文的示例通过以高效的方式对数据进行压缩以减少用于存储机器学习模型的数据的存储空间量,来增强机器学习过程。对数据进行压缩以减小数据大小会产生数个增强,这些增强包括减少从存储器淘汰和获取的量、减少用于传输经压缩的数据的带宽以及降低执行机器学习模型的存储器占用(memory footprint)。此外,由于数据被压缩,因此机器学习模型可以利用所减少的计算资源(例如,通过在较少数量的芯片上操作)执行(例如,机器学习模型可能不会基于存储器约束而被拆分为在多个芯片上操作)。

[0029] 为了实现上述内容,多个示例基于第一压缩方案对第一数据进行压缩以生成第二数据,其中,第一数据与第一机器学习模型相关联;将第二数据存储到存储器中;根据第一数据是基于第一压缩方案而被压缩,来调整查找表的第一条目以对应于第一压缩方案;在执行第一机器学习模型期间,向处理阵列的处理元件提供来自存储器的第二数据;以及在处理阵列处,基于查找表对第二数据进行解压缩以获得第一数据。第二数据(经压缩的数据)可以被存储在存储器中,以减少这些数据相对于未经压缩的数据(例如,第一数据)的存储器占用,减少计算资源和功率。这样做减少了淘汰和获取的量,以减少延迟和计算资源,同时仍保持准确性。

[0030] 也就是说,在未经压缩的数据或经低效压缩的数据上执行的现有实施方式在使用高质量的机器学习模型进行操作时是一项挑战。本文的示例涉及技术改进:增强的计算机压缩和解压缩系统。示例包括计算机能力的特定改进(例如,可以平衡性能和重复使用考虑的智能压缩方案),而不是“仅将计算机作为工具来调用的、被视为‘抽象概念’的过程”。因此,相对于现有的示例,本文的示例是增强的。

[0031] 现在转到图1,示出了计算架构100(例如,一个或多个计算设备、一个或多个服务器、一个或多个计算节点等)。出于上述性能原因(例如,更低的延迟、减少的存储器使用、减少的计算资源等),计算架构100可以采用增强的压缩和解压缩方案来将机器学习模型110的数据(例如,权重、偏置、输入数据等)存储到片上存储器120(例如,片上静态随机存取存储器(static random-access memory,SRAM))中。替代地,如果机器学习模型110的数据不能放入片上存储器120中,则数据可能必须存储在不在芯片132上的片外存储器138(例如,动态随机存取存储器)中,并且在不同的时间段期间,数据可能会被分片段带入到片上存储器120和/或处理阵列124中,从而增加了延迟、功率和传输开销。在另一替代中,机器学习模

型110可能需要在多个芯片和/或设备上操作以实现模型并行化。然后,对来自该多个芯片和/或设备的输出进行组合以生成最终输出。这样做会显著增加计算资源和功率消耗。本文的示例解决了上述问题以节省功率,因为对片外存储器138的访问将减少,并且还减少了带宽(例如,快速外围部件互连(Peripheral Component Interconnect Express,PCIE)带宽)以及计算资源(例如,在数量减少的设备和/或芯片上执行)。例如,机器学习模型110的权重可以被压缩,使得所有权重可以被同时存储到片上存储器120中,然后可以在多个处理元件中的对应处理元件中消耗时被解压缩。

[0032] 例如,可以生成机器学习模型110。可以在多个训练批次的迭代过程中,对机器学习模型110进行训练。在每次迭代期间,机器学习模型110可以基于该多个批次中的一批次来训练,并且基于机器学习模型110对该批次的操作的准确度(例如,经由反向传播过程)来更新。在训练期间,可以对机器学习模型110的数据(例如,权重和偏置)进行更新。这些数据可以被称为经解压缩的数据116。在一些示例中,数据还可以包括输入(例如,数据的特征和/或特性,机器学习模型110使用其进行预测或分类)。

[0033] 经解压缩的数据116可以是机器学习模型110的参数(例如,权重和偏置,其用于计算激活)。经解压缩的数据116可能会占用大量的存储器空间。为了减小经解压缩的数据116的大小,压缩分析器112可以对经解压缩的数据116进行分析以确定不同的压缩方案,以有效地对经解压缩的数据116进行压缩。压缩分析器112可以在机器学习模型110被生成时操作,和/或可以在用机器学习模型110进行的推理之前离线操作,以避免推理期间的延迟。可以将多个压缩方案应用于经解压缩的数据116。在一些示例中,仅一个压缩方案被应用于经解压缩的数据116,其中,该压缩方案的不同部分被存储在查找表114的不同条目中。

[0034] 详细地,在编译时,压缩分析器112(例如,软件实施方式)可以在机器学习模型110的模型创建时对经解压缩的数据116进行压缩,以生成经压缩的数据118、生成元数据134并生成查找表114。元数据134是如下信息:所述信息用于帮助对经压缩的数据118进行解压缩以获得经解压缩的数据116。如下面将解释的,压缩分析器112生成查找表114,并且可以被更新以包含与元数据134相关联的解压缩方案。对查找表114进行访问,以对经压缩的数据118进行解压缩。查找表114可以包括特定数量的(例如,512个)条目,这些条目中的每个条目都是预定量的比特(例如,21比特)。这些比特可以仅仅与非对称数字系统(asymmetric numeral system,ANS)解码算法相关,该ANS解码算法用于对经压缩的数据进行解码。

[0035] 压缩分析器112在推理时间之前对经解压缩的数据116进行分析,并确定对于给定元素大小(例如,8比特或8比特符号)最常使用哪些值(最常用值)。然后,使用数个比特对这些最常用值进行编码,该数个比特少于表示经解压缩的数据116中的最常用值的原始比特数,而最不可能被使用的值可能会占用与原始值一样多的比特,或者如果需要,可能会占用更多的比特。在数据实际用于任何类型的计算之前,压缩分析器112可以离线执行。

[0036] 经压缩的数据118(例如,经压缩的权重和偏置)、元数据134和查找表114可以与元数据134相关联地存储到片上存储器120和/或片外存储器138中。在一些示例中,片外存储器138可以在处理阵列124执行机器学习模型110之前,存储经压缩的数据118、元数据134和查找表114。当处理阵列124在推理期间执行机器学习模型110时,经压缩的数据118从片外存储器138被传输到片上存储器120。经压缩的数据118包括经压缩的权重、偏置和/或输入数据。元数据134包括如下信息:所述信息描述了在经压缩的数据118的连续存储区域中发

生的压缩。压缩和解压缩算法例如可以是非对称数字系统(ANS)。在一些示例中,元数据134是基于这样的基本假设而生成的,即仅使用ANS来对经解压缩的数据116进行压缩;并且查找表114的不同条目中的每个条目对应于经压缩的数据118的一不同部分(例如,如何对该部分进行解压缩)。例如,查找表114中的不同条目中的每个条目最终对应于要由传入的经压缩数据流产生的不同8比特符号(即,经解压缩的数据)。给定一来自经压缩的数据118的经压缩数据流,查找表114最终保存经压缩的数据与所应当对应的经解压缩的数据之间的映射,并且还保存关于如何对传入的经压缩数据流中的下一个元素进行解压缩的信息。这样做减轻了通过元数据134指示压缩算法是ANS的任何义务。在一些示例中,包括经压缩的数据118的每个压缩块可以具有相关联的4字节(B)元数据(例如,元数据134)。元数据134可以用于找到对应的经压缩的数据(例如,经压缩的第一数据和/或经压缩的第二数据)所在的最终位置,以及经压缩的数据的大小。以下是元数据的字段:

[0037] [31]:有意比特,用于指示编程是有意的。

[0038] [30:27]:左半压缩大小,4'b0000＝1 32B-beat(拍)

[0039] [26:23]:右半压缩大小,4'b0000＝1 32B-beat

[0040] [22:0]:数据偏移

[0041] 以32B对齐方式将给定经压缩的缓冲区的元数据存储在片上存储器120(例如,低延迟有效(LLS)/动态随机存取存储器(dynamic random access memory,DRAM))和/或片外存储器138的连续区域中,其中,每个32B包含8个4B元数据。在一些示例中,可以绕过片上存储器120,并且可以将经压缩的数据118、元数据134和查找表114的部分从片外存储器138传输到处理阵列124。

[0042] 在该示例中,压缩分析器112将不同的压缩方案应用于经解压缩的数据116。压缩分析器112可以确定不同的压缩方案是否对数据进行有效地压缩,并且响应于不同压缩方案中的一压缩方案满足阈值,而将该压缩方案应用于经解压缩的数据116。例如,压缩分析器112基于不同压缩方案中的第一压缩方案识别第一数据的第一部分(例如,块)被压缩了一量,并且基于该量满足存储器阈值而确定要用第一压缩方案对第一部分进行压缩。存储器阈值可以是被设置为确保所有经压缩的数据118能够同时存储在片上存储器120中的值,例如可用存储器空间总量除以经压缩的数据118的块和/或部分的数量。

[0043] 例如,片上存储器120的可用存储器空间可以被限制为特定值。经解压缩的数据116的大小可能大于该特定值,这意味着所有经解压缩的数据116不能同时被存储到片上存储器120中。因此,压缩分析器112可以搜索如下压缩方案:该压缩方案将经解压缩的数据116的大小减小到等于或小于片上存储器120的可用存储器空间以减少上述的存储器获取、功率、延迟等。在该示例中,可以确定第一压缩方案,以将经解压缩的数据116的大小减小到小于特定值。在一些示例中,压缩分析器112可以将测试压缩方案应用于经解压缩的数据116以生成测试数据,并且可以确定测试数据是否能够放入片上存储器120中。如果是,则压缩分析器112可以确定测试压缩方案被应用于经解压缩的数据116。否则,测试压缩方案将被绕过,并且将不会根据第一压缩方案对经解压缩的数据116进行压缩。

[0044] 压缩分析器112在识别要应用于经解压缩的数据116的压缩方案时,还可以考虑与查找表114有关的延迟和开销方面。处理阵列124可以基于查找表114对数据进行解压缩。查找表114将不同压缩方案与查找值相关联地存储。元数据(例如,元数据134)可以引用来自

查找表114的查找值来检索压缩方案,以对与该元数据相关联的数据进行解压缩。在一些示例中,查找表114可以仅包括一个压缩方案,其中各个条目对应于不同的值(例如,符号)。

[0045] 在一些示例中,ANS解压缩算法使用查找表114来查找所计算的9比特值并产生8比特符号。为了限制扇出问题,在该设计中,查找表114可以被复制四次,使得查找表114中的每个条目具有为8的扇出。每个条目可以是21比特,从最高有效位开始到最低有效位包括在该周期(cycle)中解压缩的4比特“nbit”值、9比特“newX”值和8比特符号。

[0046] 在每个周期中,可以执行以下操作。可以存在与在一周期中被解压缩的每个元素相关联的nbit值。每个nbit值指示每个经压缩的元素占据多少比特(例如,基于由该元素实际耗尽的比特数作为掩码来操作,如由与每个元素相关联的“nbit”值所指示的,有效值为0至9)。nbit值来自查找表114,除了第一个周期(其中,所有nbit值都是9),因为读取的第一组数字是起始查找表114索引并且查找表114索引是9比特。nbit值用于计算每个经压缩的元素的开始所在的位置。例如,元素[0]开始于比特[0],元素[1]开始于nbit[0],元素[2]开始于(nbit[0]+nbit[1]),元素[3]开始于(nbit[0]+nbit[1]+nbit[2]),依此类推。为了计算出从传入的比特流中总共消耗了多少比特,示例将所有32个4比特的nbit值相加。这将是右移位量,该右移位量用于向右移位中转缓冲区中的当前周期的内容,以便为下一个周期做好准备。一旦知道了所有元素的开始位置,就可以使用这些开始位置来对适当的比特(例如,最坏的情况是每个经压缩的元素是9比特)进行多路选择器(MUX)输出。一旦识别了与每个元素相关联的9比特值,则基于由元素实际使用的比特数来掩码9比特值,如与每个元素相关联的“nbit”值所指示的。其输出对于每个元素仍然是9比特值,其中,最高位被适当地掩码。然后,对于每个元素,将9比特掩码值添加到9比特“newX”素数。“newX”素数除了在第一个周期是0之外,可以是来自查找表114的“newX”值。因此,可以存在从查找表114读取的三个值:每个元素的未经压缩的8比特符号、要被用于下一周期的nbit值、以及要被用于下一周期的newX值。在一些示例中,查找表114可以存储多个不同的压缩方案,其中每个压缩方案具有独特的压缩细节,这些压缩细节被存储在查找表114的一组条目中。

[0047] 查找表114可能能够存储特定数量的条目(从第一条目到第N条目),各个条目包括不同的压缩方案。当新数据被加载到处理阵列124时,查找表114中的压缩方案可以被更新和/或改变,以反映新数据是如何被压缩的。例如,可以用先前未存储在查找表114中的新压缩方案,对新数据进行压缩。因此,当新数据被加载到处理阵列124时,查找表114可以被相应地更新以包括新压缩方案。当新数据被解压缩时,与新数据相关联的元数据可以引用查找表114中的合适的条目以从合适的位置检索新压缩方案和/或条目。可以基于从查找表114检索到的新压缩方案,对新数据进行解压缩。因此,基于压缩方案是否将被多次重复使用来选择压缩方案,以抵消更新查找表114和/或将查找表114加载到第一处理元件104至第N处理元件106的开销。

[0048] 例如,压缩分析器112可以预测第一压缩方案要被应用于经解压缩的数据116至少第一次数,确定该至少第一次数满足重复使用阈值(可以被选择为抵消更新查找表的开销的数量),基于该至少第一次数满足重复使用阈值而选择第一压缩方案,用第一压缩方案对经解压缩的数据116进行压缩以生成经压缩的数据118并将经解压缩的数据116的大小减小到小于或等于片上存储器120的可用存储器空间,生成元数据134以将第一压缩方案与经压缩的数据118相关联,并将经压缩的数据118与元数据134相关联地存储到片上存储器120

中。在一些示例中,经压缩的数据118可以与元数据134相关联地存储在片外存储器138中,并且然后可以在执行机器学习模型110时被提供给片上存储器120。

[0049] 压缩分析器112还可以确定压缩方案的摊销(amortization)方面。也就是说,如果压缩方案将经解压缩的数据116压缩(例如,应用多次压缩方案)到超过更新查找表114以包括该压缩方案的开销的程度,则可以选择该压缩方案。在一些示例中,可以识别基于被应用于经解压缩的数据116的压缩方案的压缩比,并将该压缩比与压缩阈值进行比较。如果压缩比满足压缩阈值,则可以认为压缩方案的益处超过更新查找表114的开销,并且压缩分析器112可以基于该压缩方案来对经解压缩的数据116进行压缩。

[0050] 在一些示例中,压缩分析器112还从延迟角度确定进一步应用压缩是否将是有益的。例如,经解压缩的数据116可以被压缩相当大的量。在一压缩点之后,进一步压缩的缺点超过益处。例如,查找表114可以被调整和/或修改以包括压缩方案。详细地,假设查找表114可以仅存储四个条目。如果使用五种压缩方案来将经解压缩的数据116压缩成经压缩的数据118,则查找表114将被调整以替换压缩方案,因为并非所有五种压缩方案可以同时被存储在查找表114中。更新和/或修改查找表114可能会增加延迟。如果仅利用这些压缩方案中的四种压缩方案就可以将经解压缩的数据116压缩成经压缩的数据118,使得所有经压缩的数据118都可以被存储在片上存储器120中,则第五压缩方案可以被视为在增加延迟的同时提供最小的存储器益处。因为这样做提供了最小的存储器优势并且增加了延迟,因此可以绕过第五压缩方案来压缩经解压缩的数据116。因此,一些示例可以仅将经解压缩的数据116压缩一量,以确保所有经压缩的数据118都可以被存储在片上存储器120中,并且当所有经压缩的数据118都能够被存储在片上存储器120中时,停止压缩和/或停止供压缩分析器112进一步压缩的分析。

[0051] 在上述中,值得注意的是,经解压缩的数据116可以被分成块(例如,32×32字节(32x 32byte)大小,具有来自一个权重矩阵和/或多个权重矩阵的int8/uint8权重的1k个元素)。给定块可以是经压缩的,也可以是未经压缩的。

[0052] 在该示例中,压缩分析器112确定利用第一压缩方案来对经解压缩的数据116进行压缩。压缩分析器112生成元数据134。元数据134基于查找表114提供用于对经压缩的数据118进行解压缩的指令和/或数据。在一些示例中,可以用多个压缩方案对经解压缩的数据116进行压缩,并且元数据134可以反映同样多的内容。

[0053] 当机器学习模型110执行推理时,经压缩的数据118和元数据134可以被存储在芯片132的片上存储器120中。例如,处理阵列124可以实施机器学习模型110以执行推理。芯片132可以在物理上具有形成在该芯片132上的处理阵列124和片上存储器120。

[0054] 在用机器学习模型110进行推理期间,处理阵列124可以接收经压缩的数据118。处理阵列124包括多个处理元件。例如,该多个处理元件中的第一处理元件104接收经压缩的第一数据,该经压缩的第一数据是经压缩的数据118的第一部分。该多个处理元件中的第N处理元件106接收经压缩的第二数据,该经压缩的第二数据是经压缩的数据118的第二部分,该第二部分不同于第一部分。第一处理元件104和第N处理元件106都接收元数据134。

[0055] 第一处理元件104的第一解压缩引擎104a基于元数据134和查找表114(例如,实时地)对经压缩的第一数据进行解压缩以生成经解压缩的第一数据。如下文将论述的,元数据134可以提供指令以基于查找表114对经压缩的第一数据进行解压缩。第一解压缩引擎104a

向第一计算元件104b提供经解压缩的第一数据。在一些示例中,第一处理元件104不是接收所有的元数据134,而是接收元数据134的、对应于经压缩的第一数据的第一部分,而不是元数据134的、不对应于经压缩的第一数据的其它部分。这样做可以减少通信带宽。

[0056] 第一处理元件104包括第一计算元件104b。第一计算元件104b可以是算术逻辑单元、加法器、乘法器等。第一计算元件104b还可以包括各种逻辑门。第一计算元件104b对经解压缩的第一数据执行运算以生成第一值,该第一值包括输出122的一部分。

[0057] 第N处理元件106的第N解压缩引擎106a基于元数据134和查找表114(例如,实时地)对经压缩的第二数据进行解压缩以生成经解压缩的第二数据。如下文将论述的,元数据134可以提供指令以基于查找表114对经压缩的第二数据进行解压缩。第N解压缩引擎106a向第N计算元件106b提供经解压缩的第二数据。在一些示例中,第N处理元件106不是接收所有的元数据134,而是接收元数据134的、对应于经压缩的第二数据的第二部分,而不是元数据134的、不对应于经压缩的第二数据的其它部分。这样做可以减少通信带宽。

[0058] 第N处理元件106包括第N计算元件106b。第N计算元件106b可以是算术逻辑单元、加法器、乘法器等。第N计算元件106b还可以包括各种逻辑门。第N计算元件106b对经解压缩的第二数据执行运算以生成第二值,该第二值包括输出122的一部分。

[0059] 注意,压缩分析器112离线地并在推理期间执行机器学习模型110之前,将经解压缩的数据116压缩成经压缩的数据118。这样做可以减少推理期间的延迟。在此类示例中,经压缩的数据118可以存储到片外存储器138中,并且然后可以在推理期间从片外存储器138被移动到片上存储器120中。

[0060] 此外,因为经压缩的数据118被存储在片外存储器138和/或片上存储器120中,因此相对于将经解压缩的数据116存储到片外存储器138和/或片上存储器120中,经压缩的数据118在片外存储器138和/或片上存储器120中占用的存储器更少。此外,减少了带宽,这是因为经压缩的数据118在压缩分析器112、片外存储器138和/或片上存储器120之间传输。此外,在第一处理元件104处对经压缩的第一数据进行解压缩以及在第N处理元件106处对经压缩的第二数据进行解压缩,以避免传输所有经解压缩的数据116,从而减少了通信带宽和计算带宽。

[0061] 在一些示例中,可以根据数个压缩方案(例如,第一压缩方案和第N压缩方案)对经解压缩的数据116的不同部分进行压缩。在此类示例中,查找表114被调整以包括第N压缩方案,并且元数据134被相应地调整。

[0062] 在一些示例中,第一处理元件104至第N处理元件106可以接收不同的查找表。例如,假设第一处理元件104正在执行与机器学习模型110(例如,第一机器学习模型)相关联的第一工作负载。此外,假设第N处理元件106正在执行与第二机器学习模型相关联的第二工作负载。与第一机器学习模型相关联的第一数据将不同于与第二机器学习模型相关联的第二数据。因此,应用于第二机器学习模型的压缩方案可以不同于应用于第一机器学习模型的压缩方案。因此,可以为第一机器学习模型和第二机器学习模型生成不同的查找表。因此,第一处理元件104可以接收第一查找表以对第一数据进行解压缩,而第N处理元件106可以接收不同于第一查找表的第二查找表(包括与第一查找表不同的压缩方案)。在其它示例中,向处理阵列124的所有处理元件提供同一查找表114。

[0063] 在一些示例中,不同的数据可以被不同地压缩。例如,可以根据第二压缩方案对其

它数据进行压缩,并相应地对其进行解压缩。其它数据可以与机器学习模型110相关联,或者可以与另一机器学习模型相关联。

[0064] 因此,本文描述的示例减少了执行机器学习模型110的延迟、计算资源和存储器内存(memory storage)。此外,查找表114可以被即时地调整,并且允许适应与机器学习模型110相关联的数据的改变。

[0065] 例如,假设经解压缩的数据116对应于机器学习模型110的第一层。在处理阵列124完成对第一层的执行之后,表示机器学习模型110的第二层的新数据可以被识别。可以基于如下的不同压缩方案来对新数据进行压缩(类似于如上所述的):这些压缩方案不同于被应用于经解压缩的数据116以生成经压缩的数据118的第一压缩方案和第N压缩方案。查找表114可以被调整,以包括用于新数据的不同压缩方案并移除第一压缩方案和第N压缩方案。此外,新数据可以与新元数据相关联。多个处理元件可以基于新元数据和经更新的查找表114来对新数据进行解压缩(与上述类似),并且可以执行与第二层相关联的计算。在一些示例中,经解压缩的数据116可以根据多个压缩方案来解压缩。

[0066] 计算架构100可以在计算设备(包括存储器和处理器)、计算系统(例如,硬件、可配置逻辑、固定功能逻辑硬件、包括用于执行的指令集的至少一个计算机可读存储介质等)等中实现。本文的示例可以在软件和/或硬件中操作。

[0067] 在一些示例中,计算架构100基于第一压缩方案对经解压缩的数据116(例如,第一数据)进行压缩以生成经压缩的数据118(例如,第二数据),其中,经解压缩的数据116与机器学习模型110(例如,第一机器学习模型)相关联。在此类示例中,计算架构100将经压缩的数据118(例如,第二数据)存储到片外存储器138和/或片上存储器120(例如,存储器)中的一者或多者中;根据经解压缩的数据116(例如,第一数据)是基于第一压缩方案而被压缩,来调整查找表114的第一条目以对应于第一压缩方案;在机器学习模型110的执行期间向处理阵列124的处理元件(例如,第一处理元件104和第N处理元件106)提供来自片外存储器138和/或片上存储器120的经压缩的数据118(例如,第二数据);以及在处理阵列124处,基于查找表114对经压缩的数据118(例如,第二数据)进行解压缩以获得第一数据。

[0068] 计算架构100还可以基于第二压缩方案对第三数据进行压缩以生成第四数据,其中,第三数据与第二机器学习模型相关联;将第四数据存储到片外存储器138和/或片上存储器120中;根据第三数据是基于第二压缩方案而被压缩,来调整查找表114的第二条目以对应于第二压缩方案;在第二机器学习模型的执行期间,向处理阵列124提供来自片外存储器138和/或片上存储器120的第三数据;以及在处理阵列124处,基于查找表114对第四数据进行解压缩以获得第三数据。在此类示例中,计算架构100用第二条目替换查找表114中的第一条目。在其它示例中,计算架构100将第二条目插入到查找表114中,使得查找表114包括第一条目和第二条目。

[0069] 在一些示例中,计算架构100基于第一压缩方案识别经解压缩的数据116的第一部分可压缩一量,并基于该量满足压缩阈值确定要用第一压缩方案对该第一部分进行压缩。

[0070] 在一些示例中,计算架构100确定,用第一压缩方案将经解压缩的数据116压缩成经压缩的数据118将经解压缩的数据116的大小减小到小于或等于片上存储器120的可用存储器空间;预测第一压缩方案要被应用至少第一次数;确定该至少第一次数满足重复使用阈值;基于该至少第一次数满足重复使用阈值,选择第一压缩方案;以及用第一压缩方案对

第一数据进行的压缩将经解压缩的数据116的大小减小到小于或等于可用存储器空间;生成元数据134以将第一压缩方案与经压缩的数据118相关联;以及将经压缩的数据118与元数据133相关联地存储到片上存储器120中。

[0071] 图2示出了用于对与机器学习模型(例如,神经网络)相关联的数据进行压缩的方法450。方法450的一个或多个方面可以被实现为计算架构100(图1)的一部分和/或结合计算架构100(图1)来实现。方法450可以在计算设备、计算系统(例如,硬件、可配置逻辑、固定功能逻辑硬件、包括用于执行的指令集的至少一个计算机可读存储介质等)中实现。

[0072] 所示的过程框452生成机器学习模型。所示的过程框454识别机器学习模型的权重和偏置。所示的过程框456识别可应用于权重和偏置以对权重和偏置进行压缩的压缩方案。所示的过程框458识别每个压缩方案的重复使用率(例如,每个压缩方案可以对多少数据块进行压缩、压缩率等)。所示的过程框460识别具有多个重复使用率中的最高重复使用率的一个或多个第一压缩方案。在一些示例中,过程框460包括基于一个或多个第一压缩方案能够对权重和偏置进行压缩使得经压缩的权重和偏置能够同时存储在片上存储器中,来选择该一个或多个第一压缩方案。所示的过程框462根据该一个或多个第一压缩方案对权重和偏置进行压缩,并生成与经压缩的权重和偏置的位置和大小以及该一个或多个第一压缩方案相对应的元数据。所示的过程框462可以包括绕过多个压缩方案中的用于对权重和偏置进行压缩的第二压缩方案。第二压缩方案可以具有如下重复使用率:该重复使用率低于该一个或多个第一压缩方案的重复使用率。当用第二压缩方案进行压缩的益处是极小的或低于阈值(例如,第二压缩方案的重复使用率低于阈值)时,避免将第二压缩方案用于压缩权重和偏置可以减少更新查找表以反映第二压缩方案的开销。所示的过程框464将经压缩的权重和偏置以及元数据存储在存储器(例如,片外存储器)中。

[0073] 图3示出了用于对机器学习数据进行压缩和解压缩的方法470。方法470的一个或多个方面可以被实现为计算架构100(图1)和/或方法450(图2)的一部分,和/或结合计算架构100(图1)和/或方法450(图2)来实现。方法470可以在计算设备、计算系统(例如,硬件、可配置逻辑、固定功能逻辑硬件、包括用于执行的指令集的至少一个计算机可读存储介质等)中实现。

[0074] 所示的过程框472确定第一机器学习模型将在处理阵列上被执行,其中,第一机器学习模型的数据被压缩并被存储在片外存储器中。所示的过程框474识别对第一机器学习模型的数据进行压缩的第一压缩方案。所示的过程框476调整查找表以包括与(例如,存储在特定槽(slot)中的)标识符相关联的第一压缩方案。所示的过程框478将查找表传送到处理阵列的处理元件。所示的过程框480将经压缩的数据从片外存储器传输到与处理阵列相关联的片上存储器。所示的过程框482基于与经压缩的数据相关联地存储的元数据以及查找表,来对经压缩的数据进行解压缩。所示的过程框484基于经解压缩的数据并利用处理阵列来执行第一机器学习模型。

[0075] 图4示出了用于在处理阵列上执行多个机器学习模型的方法490,其中对本文所描述的机器学习数据进行增强的压缩和解压缩。方法490的一个或多个方面可以被实现为计算架构100(图1)、方法450(图2)和/或方法470(图3)的一部分,和/或结合计算架构100(图1)、方法450(图2)和/或方法470(图3)来实现。方法490可以在计算设备、计算系统(例如,硬件、可配置逻辑、固定功能逻辑硬件、包括用于执行的指令集的至少一个计算机可读存储介

质等)中实现。

[0076] 所示的过程框492识别要在处理阵列上同时执行的多个机器学习模型。所示的过程框494识别与这些机器学习模型相关联的压缩方案。所示的过程框496调整查找表以包括这些压缩方案。在一些示例中,不同机器学习模型利用不同的查找表。过程框496将这些压缩方案与标识符相关联地存储。标识符可以是与要被解压缩的数据相关联的元数据中的引用。例如,过程框496包括与用于该模型的压缩方案相关联的查找表(lookup table,LUT)值(其是在执行方法490之前离线生成的),这些查找表值与压缩方案一起被适当地加载到存储器中。所示的过程框498将查找表(或多个查找表)传送到处理阵列的处理元件。在一些示例中,过程框498包括根据不同处理元件正执行的机器学习模型,向这些不同处理元件发送不同的查找表。所示的过程框500将(与机器学习模型相关联的)经压缩的数据从片外存储器传送到处理阵列的片上存储器。所示的过程框502基于与经压缩的数据相关联的元数据以及查找表来对经压缩的数据进行解压缩。所示的过程框504利用处理阵列并基于经解压缩的数据来执行该多个机器学习模型。

[0077] 图5示出了用于对机器学习模型的数据进行压缩的方法510。方法510的一个或多个方面可以被实现为如下项的一部分和/或结合如下项来实现:计算架构100(图1)、方法450(图2)、方法470(图3)和/或方法490(图4)。方法510可以在计算设备、计算系统(例如,硬件、可配置逻辑、固定功能逻辑硬件、包括用于执行的指令集的至少一个计算机可读存储介质等)中实现。

[0078] 所示的过程框512识别适用于机器学习模型的数据的相应压缩方案。所示的过程框514识别可以利用该相应压缩方案对其进行压缩的数据块。所示的过程框516将这些块的数量设置为相应压缩方案的重复使用率。所示的过程框518确定是否对所有压缩方案进行了重复使用率分析。如果不是,则所示的过程框520选择尚未分析重复使用率的另一可用压缩方案作为该相应压缩方案,并且过程框514重新执行。否则,所示的过程框522从多个重复使用率中选择具有最高重复使用率的一个或多个压缩方案,来对机器学习模型的数据进行压缩。所示的过程框524将经压缩的数据与标识如何对数据进行压缩的元数据相关联地存储。

[0079] 图6A和图6B是示出了经压缩的缓冲区的生命周期的示意图。图6A示出了经解压缩的缓冲区200如何包括未经压缩的权重矩阵。未经压缩的权重矩阵被分解成更小的32×32块。图6B示出了未经压缩的和经压缩的版本将如何在存储器中布局。经解压缩的缓冲区200、未经压缩的块的存储器布局250、以及经压缩的块的存储器布局252的一个或多个方面可以被实现为如下项的一部分和/或结合如下项来实现:计算架构100(图1)、方法450(图2)、方法470(图3)、方法490(图4)和/或方法510(图5)。方法470可以在计算设备、计算系统(例如,硬件、可配置逻辑、固定功能逻辑硬件、包括用于执行的指令集的至少一个计算机可读存储介质等)中实现。

[0080] 图6A示出了经解压缩的缓冲区200,并且图6B示出了存储器布局250。在机器模型创建时,示例(例如,基于软件和/或硬件)可以决定修剪(prune)机器学习模型的权重矩阵,或者保持权重矩阵不变,然后应用压缩(例如,非对称数字系统压缩算法)。修剪可以提供可接受的压缩比结果。

[0081] 当配置包括用于在给定模型上工作的处理阵列的半导体芯片时,主机软件可以将

经压缩的权重和/或偏置以及对应的元数据一起复制到片上存储器(例如,片上SRAM和/或低级高速缓存)中。示例还可以将数据复制到片外存储器(例如,DRAM)中。一旦已将经压缩的数据复制到片上存储器,机器学习模型就可以通过首先使用各种指令加载查找表,然后在读取经压缩的权重时利用直接存储器访问指令(称为数据获取请求)来开始运行。直接存储器访问指令可以由直接存储器访问硬件块支持。因为此类指令正在读取未经压缩的数据,因此直接存储器访问硬件块可以作为直通块来操作。直接存储器访问指令可以包含各种字段(例如,数据和元数据基地址、预获取提示等)。可以用单独的直接存储器访问指令来读取每个经压缩的块。

[0082] 在这一点上,直接存储器访问硬件块可以操作以处理和读取经压缩的块。对于给定的经压缩的缓冲区,一些块可能被压缩,而其它块可能未被压缩。然而,经压缩的缓冲区的所有块可以具有相关联的元数据(例如,4个字节),并且基于已读取到的该元数据来识别经压缩的缓冲区的块是否是未经压缩的。因此,最初可以先占用重排序缓冲区(re-order buffer,ROB)槽,但是在读取元数据并确定给定块实际上未被压缩之后,可以为给定块解除分配ROB槽。ROB具有用于如下多个请求的条目/槽:这些条目/槽用于在指令(例如,直接存储器访问指令)请求的整个生命周期中追踪该指令请求并保持与该请求相关联的数据。数据拍(data beat)无序返回,并被放置到ROB中。一旦所有拍被返回,就可以对这些拍进行解压缩。

[0083] 对于给定的经压缩的缓冲区,允许一些块被压缩同时仍允许其它块未被压缩,这意味着示例仍将对与未经压缩的块相关联的元数据进行分配,以允许更好地打包高速缓存(例如,机载存储器)中的数据。在一些示例中,所有块都将具有一些压缩量,但由于一些固有的算法开销,存储未经压缩的一些块可能会更便宜。

[0084] 经解压缩的缓冲区200包括多个32×32块。未经压缩的缓冲区占用了大量的存储器空间。图6B示出了存储器布局250和经压缩的块的存储器布局252。如图所示,存储器布局250中的数据偏移按照约每1024字节是可预测的。经压缩的块的存储器布局252中的数据偏移是不可预测的。在元数据中提供用于经压缩的块的存储器布局252的数据偏移。

[0085] 图7示出了每个块在被压缩、布局并被存储在存储器中之前以及从存储器读取之后所经历的过程300。过程300的一个或多个方面可以被实现为如下项的一部分和/或结合如下项来实现:计算架构100(图1)、方法450(图2)、方法470(图3)、方法490(图4)、方法510(图5)、经解压缩的缓冲区200(图6A)、存储器布局250(图6B)和/或经压缩的块的存储器布局252(图6B)。模型创建312包括生成32×32未经压缩的块302。然后,压缩304对32×32未经压缩的块302进行压缩。压缩304生成描述如何将32×32未经压缩的块302压缩成N字节经压缩的块306的元数据308、生成LUT,并基于32×32未经压缩的块302生成N字节经压缩的块306。N字节经压缩的块306可以被存储在存储器中。在计算引擎314上的运行期间,解压缩直接存储器访问310基于元数据308对N字节经压缩的块306进行解压缩以生成原始的32×32未经压缩的块302。

[0086] 图8例示了示出了如何对来自数据获取请求指令或元数据的字段(例如,具体地说,src_addr字段)进行分解以识别从存储器检索哪些元数据的示意图330。一个或多个方面可以被实现为如下项的一部分和/或结合如下项来实现:计算架构100(图1)、方法450(图2)、方法470(图3)、方法490(图4)、方法510(图5)、经解压缩的缓冲区200(图6A)、存储器布

局250(图6B)、经压缩的块的存储器布局252(图6B)和/或过程300(图7)。

[0087] 因为来自元数据的偏移是23比特(表示整个高速缓存空间的28比特减去用于使其32B对齐的5比特),这指示最多将只有2^23个可能的经压缩的块,并且因此块ID也将是23比特。元数据偏移的23比特特征可以通过将经压缩的块放置在高速缓存或片上存储器中的常见用例来实现;片外存储器可能能够支持的块数量远大于片上存储器中的块数量,但是可能受到23比特限制的约束。因为每个元数据是4B,以便确定与保持元数据的高速缓存行相关联的32B地址,所以可以丢弃最低或更低的3比特(addr[12:10])。剩余的20比特(addr[32:13])可以与源地址的比特[38:5]相加,以便计算出元数据的最终地址。如果数据块位于从零开始的区域,则源地址是该数据块的地址。源地址可以是存储器布局250(图6B)的左侧一组块的“数据偏移*”值。在丢弃块ID的最低3比特之前,23比特元数据地址偏移将通过比特交换机制,其中,23比特中的任何两比特可以彼此交换,如传入指令所指定的。经交换的输出被添加到元数据基地址以确定经压缩的数据的位置。

[0088] 在一些示例中,上下文与经压缩的缓冲区读取相关联。软件读取每个经压缩的缓冲区都将消耗新的上下文。在硬件中通过来自指令的基地址对上下文进行追踪。当检测到传入的直接存储器访问指令的新基地址时,这意味着新的上下文可能会被消耗。存在两个直接存储器访问上下文,每个直接存储器访问上下文都有其自己的数据基地址。这是因为数据基地址作为传入指令的一部分,并且可以被保存,因此当ROB作出数据请求时,ROB知道使用哪个数据基地址来确定最终数据地址。硬件将追踪请求在ROB中所使用的上下文。如果检测到新的数据基地址,并且存在与当前未决上下文(outstanding context)相关联的ROB槽,则传入请求将被搁置,直到上下文变为非活动。

[0089] 图9A示出了以示例列主遍历模式遍历的经压缩的缓冲区的布局350。布局350的一个或多个方面可以被实现为如下项的一部分和/或结合如下项来实现:计算架构100(图1)、方法450(图2)、方法470(图3)、方法490(图4)、方法510(图5)、经解压缩的缓冲区200(图6A)、存储器布局250(图6B)、经压缩的块的存储器布局252(图6B)、过程300(图7)和/或示意图330(图8)。假设4B元数据,从(来自指令的)实际数据偏移获得和确定元数据偏移的一种方式是将数据偏移移位10比特。这样做会为1KB的实际数据提供4B元数据的偏移。一些示例使用32B请求,并通过将地址偏移移位13来获得32B对齐的偏移。此类方法所关注的是,对于元数据的每个高速缓存行(假设用于元数据存储装置的32B高速缓存行),8个元数据被生成,并且每个元数据可以是原始矩阵和/或缓冲区的256×32区域。可能存在以列向方式将矩阵分配给PE(处理单元)的数个情况。在此类情况下,如果PE请求32×256的缓冲区,则只有12.5%的元数据信息在PE中被利用。由于增加了元数据带宽,这将导致高速缓存抖动(thrash),并且还可能导致性能问题。

[0090] 缓解上述担忧的一种方式是启用一种在计算元数据偏移时以编程方式交换地址比特的方法。此交换将取决于矩阵大小。假设1K×1K矩阵大小,如果实际数据地址偏移的比特15和比特12被交换,则示例将在32B高速缓存行中获得128×64区域的元数据。这样做可能会使得一些缓冲区在一定水平上未得到充分利用,但这将比上述提到的蛮力方式少得多。例如,一些示例可以将64×64区域分配给PE,并且对于大多数矩阵大小,最坏情况的利用率将是50%,并且仅发生在对于以列向方式分布缓冲区的PE。比特交换可以是可编程的,并且对于大多数情况,交换两个比特可以提供操作数据。对于每块2比特的情况,示例可能

不会实现上述内容,因为元数据已经显著减少。

[0091] 为了使比特交换特征运行,示例应该知道可以在运行期间启用该特征,并将元数据正确存储在存储器中。这样做要么是首先生成元数据的编译时的软件,要么是在将元数据复制到片外存储器中时的固件。如果已知经压缩的缓冲区将以列向方式被遍历,则可以在运行时启用该特征,否则,如果经压缩的缓冲区以行向方式被遍历,则可以通过将两个交换字段设置为同一值来禁用该特征(例如,比特交换)。

[0092] 在图9A中,启用交换比特的遍历的示例概述如下:访问块0(缺失)、访问块1(命中)、访问块8(命中)、访问块9(命中)、访问块16(缺失)、访问块17(命中)、访问块24(命中)和访问块25(命中)。

[0093] 图9B示出了在没有启用交换比特的情况下将元数据存储在存储器中的布局370。布局370的一个或多个方面可以被实现为如下项的一部分和/或结合如下项来实现:计算架构100(图1)、方法450(图2)、方法470(图3)、方法490(图4)、方法510(图5)、经解压缩的缓冲区200(图6A)、存储器布局250(图6B)、经压缩的块的存储器布局252(图6B)、过程300(图7)、示意图330(图8)和/或布局350(图9A)。在本示例中,利用8个元数据示出了32个块。在图9B中,没有交换比特的示例遍历包括访问块0(缺失)、访问块1(命中)、访问块8(缺失)、访问块9(命中)、访问块16(缺失)、访问块17(命中)、访问块24(缺失)和访问块25(命中)。

[0094] 图9C示出了启用交换比特的存储器中的元数据的布局390。在本示例中,利用8个元数据示出了32个块。布局390的一个或多个方面可以被实现为以下项的一部分和/或结合以下项来实现:计算架构100(图1)、方法450(图2)、方法470(图3)、方法490(图4)、方法510(图5)、经解压缩的缓冲区200(图6A)、存储器布局250(图6B)、经压缩的块的存储器布局252(图6B)、过程300(图7)、示意图330(图8)、布局350(图9A)和/或布局370(图9B)。

[0095] 图10示出了解压缩架构420。解压缩引擎422包括与解压缩相关的逻辑。该逻辑包括到ROB的用于接受新请求的接口、到ROB随机存取存储器(random access memory,RAM)的用于发送读取请求的接口、到ROB RAM的用于存储传入的读取数据的接口、ANS解压缩逻辑以及最后的输出数据接口。解压缩架构420的一个或多个方面可以被实现为如下项的一部分和/或结合如下项来实现:计算架构100(图1)、方法450(图2)、方法470(图3)、方法490(图4)、方法510(图5)、经解压缩的缓冲区200(图6A)、存储器布局250(图6B)、经压缩的块的存储器布局252(图6B)、过程300(图7)、示意图330(图8)、布局350(图9A)、布局370(图9B)和/或布局390(图9C)。

[0096] ROB控制逻辑428将通告最老的ROB槽以使所有拍返回。RAM读取逻辑424将接受来自ROB控制逻辑428的解压缩请求,并开始向给定槽的对应RAM发出读取请求。读取的数据将被放置到触发级(flop stage)426中。因为传入数据可能在被ANS 430使用之前被左移位,因此触发级426为该移位提供完整的周期,因为在RAM读取之后直接移位可能是困难的。触发级426是单触发级还是双触发级(例如,具有寄存就绪)将取决于时序。

[0097] 然后,将数据从触发级426馈送到ANS 430,其中,根据前一周期中消耗的数据量对数据进行左移位。在ANS块已经产生了32个8比特元素之后,32个8比特元素将被输出到顶层直接存储器访问块,其中,必须根据从FI(文件标识符)432返回的数据(例如,在未经压缩的数据返回的情况下)对32个8比特元素进行仲裁。因此,解压缩引擎422的输出将是触发级434以能够正确处理搁置。

[0098] 图11示出了高级ANS解压缩流水线510。ANS核心的逻辑被认为是核心ANS解压缩逻辑。ANS解压缩流水线510可以具有4个周期的延迟(例如,从SRAM的输入开始到未经压缩的输出)并且可以保持32B/周期的吞吐量。高级ANS解压缩流水线510的一个或多个方面被实现为如下项的一部分和/或结合如下项来实现:计算架构100(图1)、方法450(图2)、方法470(图3)、方法490(图4)、方法510(图5)、经解压缩的缓冲区200(图6A)、存储器布局250(图6B)、经压缩的块的存储器布局252(图6B)、过程300(图7)、示意图330(图8)、布局350(图9A)、布局370(图9B)、布局390(图9C)和/或解压缩架构420(图10)。

[0099] 系统概述

[0100] 图12示出了与社交网络系统相关联的示例网络环境600。网络环境600可以实现已经论述过的以下项的一个或多个方面:计算架构100(图1)、方法450(图2)、方法470(图3)、方法490(图4)、方法510(图5)、经解压缩的缓冲区200(图6A)、存储器布局250(图6B)、经压缩的块的存储器布局252(图6B)、过程300(图7)、示意图330(图8)、布局350(图9A)、布局370(图9B)、布局390(图9C)、解压缩架构420(图10)和/或高级ANS解压缩流水线510(图11)。

[0101] 网络环境600包括客户端系统630、社交网络系统660和第三方系统670,该客户端系统、社交网络系统和第三方系统通过网络610彼此连接。尽管图12示出了客户端系统630、社交网络系统660、第三方系统670和网络610的特定布置,但是本公开考虑了客户端系统630、社交网络系统660、第三方系统670和网络610的任何合适的布置。作为示例而非限制,客户端系统630、社交网络系统660和第三方系统670中的两者或更多者可以绕过网络610而彼此直接连接。作为另一示例,客户端系统630、社交网络系统660和第三方系统670中的两者或更多者可以在物理上或逻辑上全部或部分地彼此位于同一位置。此外,尽管图12示出了特定数量的客户端系统630、社交网络系统660、第三方系统670和网络610,但是本公开考虑了任何合适数量的客户端系统630、社交网络系统660、第三方系统670和网络610。作为示例而非限制,网络环境600可以包括多个客户端系统630、多个社交网络系统660、多个第三方系统670和多个网络610。

[0102] 本公开考虑了任何合适的网络610。作为示例而非限制,网络610的一个或多个部分可以包括自组网、内联网、外联网、虚拟专用网(virtual private network,VPN)、局域网(local area network,LAN)、无线LAN(wireless LAN,WLAN)、广域网(wide area network,WAN)、无线WAN(wireless WAN,WWAN)、城域网(metropolitan area network,MAN)、互联网的一部分、公共交换电话网(public switched telephone network,PSTN)的一部分、蜂窝电话网、或这些网络中的两者或更多者的组合。网络610可以包括一个或多个网络610。

[0103] 多条链路650可以将客户端系统630、社交网络系统660和第三方系统670连接到通信网络610或者使客户端系统630、社交网络系统660和第三方系统670彼此连接。本公开考虑了任何合适的链路650。在特定示例中,一条或多条链路650包括一条或多条有线(例如,数字用户线路(digital subscriber line,DSL)或电缆数据服务接口规范(data over cable service interface specification,DOCSIS))链路、一条或多条无线(例如,Wi-Fi或全球微波互联接入(worldwide interoperability for microwave access,WiMAX))链路、或一个或多个光学(例如,同步光网络(synchronous optical network,SONET)或同步数字体系(synchronous digital hierarchy,SDH))链路。在特定示例中,一条或多条链路650各自包括自组网、内联网、外联网、VPN、LAN、WLAN、WAN、WWAN、MAN、互联网的一部分、PSTN

的一部分、基于蜂窝技术的网络、基于卫星通信技术的网络、另一条链路650、或两条或更多条此类链路650的组合。在整个网络环境600中,多条链路650可能不一定是相同的。一条或多条第一链路650在一个或多个方面可以不同于一条或多条第二链路650。

[0104] 在特定示例中,客户端系统630可以是如下的电子设备:该电子设备包括硬件、软件、或嵌入式逻辑部件或两个或更多个此类部件的组合,并且能够执行客户端系统630所实现或支持的适当功能。作为示例而非限制,客户端系统630可以包括计算机系统,该计算机系统例如为,台式计算机、笔记本或膝上型计算机、上网本、平板计算机、电子书阅读器、GPS设备、摄像头、个人数字助理(personal digital assistant,PDA)、手持电子设备、蜂窝电话、智能电话、增强/虚拟现实设备、其它合适的电子设备或它们的任何合适的组合。本公开考虑了任何合适的客户端系统630。客户端系统630可以使客户端系统630处的网络用户能够访问网络610。客户端系统630可以使其用户能够与其它客户端系统630处的其他用户进行通信。

[0105] 在特定示例中,客户端系统630可以包括网络浏览器632(例如微软IE浏览器(MICROSOFT INTERNET EXPLORER)、谷歌浏览器(GOOGLE CHROME)或摩斯拉火狐浏览器(MOZILLA FIREFOX)),并且可以具有一个或多个附加件(add-ons)、插件或其它扩展件(例如工具栏(TOOLBAR)或雅虎工具栏(YAHOO TOOLBAR))。客户端系统630处的用户可以输入统一资源定位符(Uniform Resource Locator,URL)或将网络浏览器632定向到特定服务器(例如,服务器662或与第三方系统670相关联的服务器)的其它地址,并且网络浏览器632可以生成超文本传输协议(Hyper Text Transfer Protocol,HTTP)请求并将该HTTP请求传送到服务器。服务器可以接受该HTTP请求,并向客户端系统630传送响应于该HTTP请求的一个或多个超文本标记语言(Hyper Text Markup Language,HTML)文件。客户端系统630可以基于来自服务器的HTML文件来渲染网页以呈现给用户。本公开考虑了任何合适的网页文件。作为示例而非限制,网页可以根据特定期望从HTML文件、可扩展超文本标记语言(Extensible Hyper Text Markup Language,XHTML)文件或可扩展标记语言(Extensible Markup Language,XML)文件渲染。这些页面还可以执行脚本,这些脚本例如但不限于用JAVA脚本(JAVASCRIPT)、JAVA、微软银光(MICROSOFT SILVERLIGHT)、标记语言和脚本的组合(例如,AJAX(异步JAVASCRIPT和XML))等编写的那些脚本。在此,在适当的情况下,对网页的引用包含一个或多个对应的网页文件(浏览器可以使用该网页文件来渲染网页),反之亦然。

[0106] 在特定示例中,社交网络系统660可以是可托管在线社交网络的网络可寻址计算系统。社交网络系统660可以生成、存储、接收和发送社交网络数据,例如用户简档数据、概念简档数据、社交图谱信息或与在线社交网络相关的其它合适的数据。社交网络系统660可以由网络环境600的其它部件直接地访问或通过网络610访问。作为示例而非限制,客户端系统630可以使用网络浏览器632或与社交网络系统660相关联的本地应用(例如,移动社交网络应用、消息收发应用、另一合适的应用或其任何组合)来直接访问社交网络系统660或经由网络610访问社交网络系统660。在特定示例中,社交网络系统660可以包括一个或多个服务器662。每个服务器662可以是单一服务器或跨多个计算机或多个数据中心的分布式服务器。服务器662可以是多种多样的,例如但不限于,网络服务器、新闻服务器、邮件服务器、消息服务器、广告服务器、文件服务器、应用服务器、交换服务器、数据库服务器、代理服务

器、适合于执行本文中所描述的功能或过程的另一服务器、或它们的任何组合。在特定示例中,每个服务器662可以包括用于执行由服务器662实现或支持的适当功能的硬件、软件、或嵌入式逻辑部件、或者两个或更多个此类部件的组合。在特定示例中,社交网络系统660可以包括一个或多个数据存储区664。数据存储区664可以用于存储各种类型的信息。在特定示例中,可以根据特定的数据结构来组织存储在数据存储区664中的信息。在特定示例中,每个数据存储区664可以是关系数据库、列式数据库、关联数据库(correlation database)或其它合适的数据库。尽管本公开描述或示出了特定类型的数据库,但是本公开考虑了任何合适类型的数据库。特定实施例可以提供多个接口,该多个接口使客户端系统630、社交网络系统660或第三方系统670能够对存储在数据存储区664中的信息进行管理、检索、修改、添加或删除。

[0107] 在特定示例中,社交网络系统660可以在一个或多个数据存储区664中存储一个或多个社交图谱。在特定示例中,社交图谱可以包括多个节点——这些节点可以包括多个用户节点(每个用户节点对应于一特定用户)或多个概念节点(每个概念节点对应于一特定概念)——以及连接这些节点的多条边。社交网络系统660可以向在线社交网络的用户提供与其他用户通信和交互的能力。在特定示例中,用户可以经由社交网络系统660加入在线社交网络,然后向社交网络系统660中他们想要联系的一些其他用户添加联系(例如,关系)。本文中,术语"朋友"可以指社交网络系统660中用户已经经由社交网络系统660与其形成了联系、关联或关系的任何其他用户。

[0108] 在特定示例中,社交网络系统660可以向用户提供在社交网络系统660支持的各种类型的项目或对象上采取行动的能力。作为示例而非限制,这些项目和对象可以包括社交网络系统660的用户可能所属的群组或社交网络、用户可能感兴趣的事件或日历条目、用户可能使用的基于计算机的应用程序、允许用户经由服务购买或出售项目的交易、用户可能执行的与广告的交互、或其它合适的项目或对象。用户可以与能够在社交网络系统660中体现的任何事物交互、或者与能够由第三方系统670的外部系统体现的任何事物进行交互,该第三方系统与社交网络系统660分离,并经由网络610耦合到社交网络系统660。

[0109] 在特定示例中,社交网络系统660可以能够连接各种实体。作为示例而非限制,社交网络系统660可以使用户能够彼此交互以及从第三方系统670或其他实体接收内容,或者允许用户通过应用程序编程接口(application programming interface,API)或其它通信通道与这些实体交互。

[0110] 在特定示例中,第三方系统670可以包括一种或多种类型的服务器、一个或多个数据存储区、一个或多个接口(包括但不限于API)、一个或多个网络服务、一个或多个内容源、一个或多个网络、或任何其它合适的部件(例如,服务器可以与这些部件通信)。可以由与操作社交网络系统660的实体不同的实体来操作第三方系统670。然而,在特定示例中,社交网络系统660和第三方系统670可以彼此协同运行,以向社交网络系统660的用户或第三方系统670提供社交网络服务。在这个意义上,社交网络系统660可以提供平台或主干网络,其它系统(例如,第三方系统670)可以使用该平台或主干网络来向互联网上的用户提供社交网络服务和功能。

[0111] 在特定示例中,第三方系统670可以包括第三方内容对象提供者。第三方内容对象提供者可以包括内容对象的一个或多个源,所述内容对象可以被传送到客户端系统630。作

为示例而非限制,内容对象可以包括关于用户感兴趣的事物或活动的信息,这些信息例如为,电影放映时间、电影评论、餐厅评论、餐厅菜单、产品信息和评论、或其它合适的信息。作为另一示例而非限制,内容对象可以包括激励内容对象,例如优惠券、折扣票、礼品券、或其它合适的激励对象。

[0112] 在特定示例中,社交网络系统660还包括用户生成的内容对象,这可以增强用户与社交网络系统660的交互。用户生成的内容可以包括用户可以向社交网络系统660添加、上传、发送或"发布"的任何内容。作为示例而非限制,用户从客户端系统630向社交网络系统660传送帖子(post)。帖子可以包括诸如状态更新或其它文本数据、位置信息、照片、视频、链接、音乐或其它类似数据或媒体等数据。也可以由第三方通过诸如新闻推送或流等"通信通道"将内容添加到社交网络系统660。

[0113] 在特定示例中,社交网络系统660可以包括各种服务器、子系统、程序、模块、日志、和数据存储区。在特定示例中,社交网络系统660可以包括以下各项中的一项或多项:网络服务器、动作记录器(action logger)、API请求服务器、相关性和排名引擎、内容对象分类器、通知控制器、动作日志、第三方内容对象暴露日志、推理模块、授权/隐私服务器、搜索模块、广告定向模块、用户界面模块、用户简档存储区、联系存储区、第三方内容存储区或位置存储区。社交网络系统660还可以包括合适的组件,这些合适的组件例如为,网络界面、安全机制、负载平衡器、故障转移服务器、管理和网络操作控制台、其它合适的组件、或它们的任何合适的组合。在特定示例中,社交网络系统660可以包括用于存储用户简档的一个或多个用户简档存储区。用户简档可以包括例如生物信息、人口统计信息、行为信息、社交信息、或其它类型的描述性信息(例如,工作经历、教育历史、爱好或偏好、兴趣、密切关系、或位置)。兴趣信息可以包括与一个或多个类别相关的兴趣。类别可以是一般的,也可以是具体的。作为示例而非限制,如果用户"喜欢"关于一鞋品牌的物品,则该类别可以是该品牌,或者可以是一般类别"鞋子"或"服装"。联系存储区可用于存储关于用户的联系信息。联系信息可以指示如下用户:所述用户具有相似或共同的工作经历、群组成员资格、爱好、教育历史,或者以任何方式相关或共享共同的属性。联系信息还可以包括不同用户与内容(内部和外部两者)之间的用户定义的联系。网络服务器可以用于经由网络610将社交网络系统660链接至一个或多个客户端系统630或一个或多个第三方系统670。网络服务器可以包括邮件服务器或用于在社交网络系统660与一个或多个客户端系统630之间接收和发送消息的其它消息传递功能。API请求服务器可以允许第三方系统670通过调用一个或多个API来从社交网络系统660访问信息。动作记录器可以用于接收来自网络服务器的关于用户在社交网络系统660上或在社交网络系统660外的动作的通信。结合动作日志,可以维护关于用户暴露于第三方内容对象的第三方内容对象日志。通知控制器可以向客户端系统630提供关于内容对象的信息。可以将信息作为通知推送到客户端系统630,或者可以从客户端系统630取出响应于从客户端系统630接收的请求的信息。授权服务器可以用于实施社交网络系统660的用户的一个或多个隐私设置。用户的隐私设置确定了可以如何共享与用户相关联的特定信息。授权服务器可以允许用户例如通过设置适当的隐私设置来选择加入或退出社交网络系统660记录他们的动作或与其它系统(例如,第三方系统670)共享他们的动作。第三方内容对象存储区可以用于存储从第三方(例如,第三方系统670)接收的内容对象。位置存储区可以用于存储从与用户相关联的客户端系统630接收的位置信息。广告定价模块可以组合社

交信息、当前时间、位置信息或其它合适的信息,以通知的形式向用户提供相关广告。

[0114] 社交图谱

[0115] 图13示出了示例社交图谱700。在一些示例中,已经讨论的计算架构100(图1)、方法450(图2)、方法470(图3)、方法490(图4)、方法510(图5)、经解压缩的缓冲区200(图6A)、存储器布局250(图6B)、经压缩的块的存储器布局252(图6B)、过程300(图7)、示意图330(图8)、布局350(图9A)、布局370(图9B)、布局390(图9C)、解压缩架构420(图10)和/或高级ANS解压缩流水线510(图11)可以访问社交图谱700以实现一个或多个方面。在特定示例中,社交网络系统660可以将一个或多个社交图谱700存储在一个或多个数据存储区中。在特定示例中,社交图谱700可以包括多个节点——这些节点可以包括多个用户节点702或多个概念节点704——以及连接这些节点的多条边706。每个节点可以与唯一的实体(即,用户或概念)相关联,每个实体可以具有唯一的标识符(identifier,ID),例如唯一的编号或用户名。出于教导目的,以二维可视化图谱表示形式示出了图13所示的示例社交图谱700。在特定示例中,社交网络系统660、客户端系统630或第三方系统670可以访问社交图谱700和相关社交图谱信息,以用于合适的应用。社交图谱700的节点和边例如可以作为数据对象存储在数据存储区(例如,社交图谱数据库)中。这种数据存储区可以包括社交图谱700的节点或边的一个或多个可搜索或可查询的索引。

[0116] 在特定示例中,用户节点702可以对应于社交网络系统660的用户。作为示例而非限制,用户可以是与社交网络系统660交互或通信、或者通过社交网络系统660交互或通信的个体(人类用户)、实体(例如,企业、公司或第三方应用程序)或群组(例如,个体群组或实体群组)。在特定示例中,当用户向社交网络系统660注册账户时,社交网络系统660可以创建与该用户相对应的用户节点702,并将该用户节点702存储在一个或多个数据存储区中。在适当的情况下,本文中所描述的用户和用户节点702可以指已注册用户和与已注册用户相关联的用户节点702。附加地或者作为替代,在适当的情况下,本文中所描述的用户和用户节点702可以指尚未向社交网络系统660进行注册的用户。在特定示例中,用户节点702可以与由用户提供的信息或由各种系统(其包括社交网络系统660)收集的信息相关联。作为示例而非限制,用户可以提供他或她的姓名、个人照片、联系信息、出生日期、性别、婚姻状况、家庭状况、职业、教育背景、偏好、兴趣、或其它人口统计信息。在特定示例中,用户节点702可以与一个或多个数据对象相关联,该一个或多个数据对象对应于与用户相关联的信息。在特定示例中,用户节点702可以对应于一个或多个网页。

[0117] 在特定示例中,概念节点704可以对应于概念。作为示例而非限制,概念可以对应于地点(例如,电影院、餐厅、地标、或城市);网站(例如,与社交网络系统660相关联的网站、或与网络应用服务器相关联的第三方网站);实体(例如,个体、公司、群组、运动队、或名人);可位于社交网络系统660内或外部服务器(例如,网络应用服务器)上的资源(例如,音频文件、视频文件、数码照片、文本文件、结构化文档、或应用);不动产或知识产权(例如,雕塑、绘画、电影、游戏、歌曲、创意、照片、或书面作品);游戏;活动;想法或理论;增强/虚拟现实环境中的对象;另一合适的概念;或者两个或更多个这种概念。概念节点704可以与由用户提供的概念的信息或由各种系统(其包括社交网络系统660)收集的信息相关联。作为示例而非限制,概念的信息可以包括名称或标题;一个或多个图像(例如,书籍的封面图像);位置(例如,地址或地理位置);网站(其可以与URL相关联);联系信息(例如,电话号码或电

子邮件地址);其它合适的概念信息;或者这些信息的任何合适的组合。在特定示例中,概念节点704可以与一个或多个数据对象相关联,该一个或多个数据对象对应于与概念节点704相关联的信息。在特定示例中,概念节点704可以对应于一个或多个网页。

[0118] 在特定示例中,社交图谱700中的节点可以表示网页(其可以被称为"简档页面")或由网页表示。简档页面可以由社交网络系统660托管或访问。简档页面也可以被托管在与第三方系统670相关联的第三方网站上。作为示例而非限制,对应于特定外部网页的简档页面可以是特定的外部网页,并且该简档页面可以对应于特定概念节点704。简档页面可以被其他用户的所有子集或选定子集查看。作为示例而非限制,用户节点702可以具有对应的用户简档页面,在该对应的用户简档页面中,对应的用户可以添加内容、做出声明、或以其它方式表达他或她自己。作为另一示例而非限制,概念节点704可以具有对应的概念简档页面,在该对应的概念简档页面中,一个或多个用户可以添加内容、做出声明、或表达他们自己,特别是在对应于概念节点704的概念方面。

[0119] 在特定示例中,概念节点704可以表示由第三方系统670托管的第三方网页或资源。除了其它元素之外,第三方网页或资源可以包括代表动作或活动的内容、可选图标或其它图标、或其它可交互的对象(其例如可以用JavaScript、AJAX或PHP代码实现)。作为示例而非限制,第三方网页可以包括可选图标,例如"喜欢"、"签到"、"吃"、"推荐"、或其它合适的动作或活动。查看第三方网页的用户可以通过选择这些图标中的一个图标(例如,"签到")来执行动作,从而使得客户端系统630向社交网络系统660发送指示用户动作的消息。响应于该消息,社交网络系统660可以在对应于用户的用户节点702和对应于第三方网页或资源的概念节点704之间创建边(例如,签到型边),并将边706存储在一个或多个数据存储区中。

[0120] 在特定示例中,社交图谱700中的一对节点可以通过一条或多条边706彼此连接。连接一对节点的边706可以表示该一对节点之间的关系。在特定示例中,边706可以包括或表示与一对节点之间的关系相对应的一个或多个数据对象或属性。作为示例而非限制,第一用户可以指示第二用户是第一用户的"朋友"。响应于该指示,社交网络系统660可以向第二用户发送"朋友请求"。如果第二用户确认该"朋友请求",则社交网络系统660可以在社交图谱700中创建将第一用户的用户节点702连接到第二用户的用户节点702的边706,并将边706作为社交图谱信息存储在一个或多个数据存储区664中。在图13的示例中,社交图谱700包括指示用户"A"的用户节点702和用户"B"的用户节点702之间的朋友关系的边706、以及指示用户"C"的用户节点702和用户"B"的用户节点702之间的朋友关系的边。尽管本公开描述或示出了连接特定用户节点702的具有特定属性的特定边706,但是本公开考虑了连接用户节点702的具有任何合适属性的任何合适的边706。作为示例而非限制,边706可以表示朋友关系、家庭关系、商业或雇佣关系、粉丝关系(例如,包括喜欢等)、关注者关系、访客关系(例如,包括访问、查看、签到、分享等)、订户关系、上级/下级关系、互惠关系、非互惠关系、另一合适类型的关系、或两个或更多个这种关系。此外,尽管本公开总体上将节点描述为被连接,但是本公开也将用户或概念描述为被连接。在本文中,在适当的情况下,对正被连接的用户或概念的引用可以指与社交图谱700中通过一条或多条边706连接的那些用户或概念相对应的节点。分别由两个节点表示的两个对象之间的分离度是社交图谱700中连接这两个节点的最短路径中的边的数目。作为示例而非限制,在社交图谱700中,用户"C"的用户

节点702通过多个路径连接到用户“A”的用户节点702,该多个路径包括例如直接通过用户“B”的用户节点702的第一路径、通过公司“爱克母(Acme)”的概念节点704和用户“D”的用户节点702的第二路径、以及通过表示学校“斯坦福大学”、用户“G”、公司“爱克母”和用户“D”的各用户节点702和概念节点704的第三路径。因为连接用户“C”的对应节点和用户“A”的对应节点的最短路径(即,第一路径)包括两个边706,所以用户“C”和用户“A”的分离度为2。

[0121] 在特定示例中,用户节点702与概念节点704之间的边706可以表示由与用户节点702相关联的用户针对与概念节点704相关联的概念而执行的特定动作或活动。作为示例而非限制,如图13所示,用户可以“喜欢”、“定期去”、“播放”、“收听”、“烹饪”、“工作在”或“观看”概念,上述项中的每一项可以对应于边类型或边子类型。对应于概念节点704的概念简档页面例如可以包括可选的“签到”图标(例如,可点击的“签到”图标)或可选的“添加到收藏夹”图标。类似地,在用户点击这些图标之后,社交网络系统660可以响应于与相应动作相对应的用户动作,来创建“收藏”边或“签到”边。作为另一个示例而非限制,用户(用户“C”)可以使用特定的应用程序(声破天(SPOTIFY),其为在线音乐应用程序)来收听特定的歌曲(“想象(Imagine)”)。在这种情况下,社交网络系统660可以在与用户对应的用户节点702和与歌曲和应用程序对应的概念节点704之间创建“收听”边706和“使用”边(如图13所示),以指示用户收听了歌曲并使用了应用程序。此外,社交网络系统660可以在与歌曲和应用程序对应的概念节点704之间创建“播放”边706(如图13所示),以指示该特定歌曲由该特定应用程序播放。在这种情况下,“播放”边706对应于外部应用程序(SPOTIFY)对外部音频文件(歌曲“想象”)执行的动作。尽管本公开描述了连接用户节点702和概念节点704的具有特定属性的特定边706,但是本公开考虑了连接用户节点702和概念节点704的具有任何合适属性的任何合适的边706。此外,尽管本公开描述了在用户节点702和概念节点704之间的边表示单个关系,但是本公开考虑了用户节点702和概念节点704之间的边表示一个或多个关系。作为示例而非限制,边706可以表示用户喜欢特定概念和用户已经使用过特定概念这两者。替代地,另一条边706可以表示用户节点702和概念节点704之间(如图13所示,用户“E”的用户节点702与“SPOTIFY”的概念节点704之间)的每种类型的关系(或多个单独的关系)。

[0122] 在特定示例中,社交网络系统660可以在社交图谱700中的用户节点702和概念节点704之间创建边706。作为示例而非限制,(例如通过使用由用户的客户端系统630托管的网络浏览器或专用应用程序)查看概念简档页面的用户可以通过点击或选择“喜欢”图标,来指示他或她喜欢由概念节点704表示的概念,这可以使得用户的客户端系统630向社交网络系统660发送指示用户喜欢与概念简档页面相关联的概念的消息。响应于该消息,社交网络系统660可以在与用户相关联的用户节点702和概念节点704之间创建边706,如由用户节点和概念节点704之间的“喜欢”边706所示。在特定示例中,社交网络系统660可以将边706存储在一个或多个数据存储区中。在特定示例中,可以由社交网络系统660响应于特定用户动作而自动形成边706。作为示例而非限制,如果第一用户上传了图片、观看了电影或收听了歌曲,则可以在对应于第一用户的用户节点702和对应于这些概念的概念节点704之间形成边706。尽管本公开描述了以特定方式形成特定边706,但是本公开考虑了以任何合适的方式形成任何合适的边706。

[0123] 社交图谱亲密度与系数

[0124] 在特定示例中,社交网络系统660可以确定各种社交图谱实体彼此间的社交图谱

亲密度(affinity)(在本文中其可以称为“亲密度”)。亲密度可以表示与在线社交网络相关联的特定对象之间的关系强度或感兴趣程度,这些特定对象例如是用户、概念、内容、动作、广告、与在线社交网络相关联的其它对象、或以上对象的任何合适的组合。还可以确定关于与第三方系统670或其它合适的系统相关联的对象的亲密度。可以建立每个用户、每个主题或每种类型的内容的社交图谱实体的总亲密度。总亲密度可以基于对与社交图谱实体相关联的动作或关系的持续监测而改变。尽管本公开描述了以特定方式确定特定的亲密度,但是本公开考虑了以任何合适的方式确定任何合适的亲密度。

[0125] 在特定示例中,社交网络系统660可以使用亲密度系数(在本文中可以称为“系数”)来测量或量化社交图谱亲密度。该系数可以表示或量化与在线社交网络相关联的特定对象之间的关系强度。该系数还可以表示这种概率或函数:该概率或函数基于用户对特定动作的兴趣来测量用户将执行该动作的预测概率。以这种方式,可以基于用户的先前动作来预测用户的未来动作,其中,可以至少部分地基于用户动作的历史来计算该系数。可以使用系数来预测任意数量的动作,这些动作可以在在线社交网络的内部或外部。作为示例而非限制,这些动作可以包括:各种类型的通信(例如,发送消息、发布内容或对内容评论);各种类型的观察动作(例如,访问或查看简档页面、媒体或其它合适的内容);关于两个或更多个社交图谱实体的各种类型的一致性信息(例如,在同一群组中、在同一照片中被标记、在同一位置处签到或参加同一活动);或其它适当的动作。尽管本公开描述了以特定方式测量亲密度,但本公开考虑了以任何合适的方式测量亲密度。

[0126] 在特定示例中,社交网络系统660可以使用各种因素来计算系数。这些因素例如可以包括用户动作、对象之间的关系类型、位置信息、其它合适的因素、或以上因素的任意组合。在特定示例中,当计算系数时,可以对不同的因素施加不同的权重。每个因素的权重可以是静态的、或者这些权重可以例如根据用户、关系类型、动作类型、用户位置等改变。可以根据这些因素的权重对这些因素的等级进行组合,以确定用户的总系数。作为一示例而非限制,可以为特定用户动作分配等级和权重两者,同时为与该特定用户动作相关联的关系分配等级和相关权重(例如,使得权重总和为100%)。为了计算用户对于特定对象的系数,分配给用户动作的等级例如可以包括总体系数的60%,而分配给用户与对象之间的关系的等级可以包括总体系数的40%。在特定示例中,当确定用于计算系数的各种因素的权重时,社交网络系统660可以考虑各种变量,这些变量例如为,信息被访问以来的时间、衰减因子、访问频率、与信息的关系或与其信息被访问的对象的关系、与连接到对象的社交图谱实体的关系、用户动作的短期平均值或长期平均值、用户反馈、其它合适的变量、或它们的任意组合。作为示例而非限制,系数可以包括衰减因子,该衰减因子使特定动作提供的信号强度随时间衰减,使得在计算系数时,越早的动作越相关。可以基于对系数所基于的动作的持续追踪,来持续更新等级和权重。可以采用任何类型的过程或算法,来对每个因素的等级以及分配给这些因素的权重进行分配、组合、平均等。在特定示例中,社交网络系统660可以使用根据历史动作和过去的用户响应而训练的机器学习算法、或者通过将用户暴露于各种选项并且测量响应而从用户获得的数据来确定系数。尽管本公开描述了以特定方式计算系数,但是本公开考虑了以任何合适的方式计算系数。

[0127] 在特定示例中,社交网络系统660可以基于用户的动作来计算系数。社交网络系统660可以监控在线社交网络上、第三方系统670上、其它合适的系统上或它们的任意组合上

的这些动作。可以跟踪或监控任何合适类型的用户动作。典型的用户动作包括查看简档页面、创建或发布内容、与内容交互、加上标记或者在图像内加上标记、加入群组、列出和确认活动的出席、在位置处签到、喜欢特定页面、创建页面以及执行促进社交动作的其它任务。在特定示例中，社交网络系统660可以基于用户对特定类型内容的动作来计算系数。该内容可以与在线社交网络、第三方系统670或另一合适的系统相关联。该内容可以包括用户、简档页面、帖子、新闻内容、头条、即时消息、聊天室对话、电子邮件、广告、图片、视频、音乐、其它合适的对象、或以上内容的任意组合。社交网络系统660可以分析用户的动作以确定多个动作中的一个或多个动作是否指示对主题、内容、其他用户等的亲密度。作为示例而非限制，如果用户频繁发布与“咖啡”或其变型相关的内容，则社交网络系统660可以确定该用户相对于概念“咖啡”具有较高的系数。可以为特定动作或特定类型的动作分配比其它动作更高的权重和/或等级，这可能影响所计算的总系数。作为示例而非限制，如果第一用户向第二用户发送电子邮件，则该动作的权重或等级可能高于在第一用户仅查看第二用户的用户简档页面的情况下的权重或等级。

[0128] 在特定示例中，社交网络系统660可以基于特定对象之间的关系类型来计算系数。参考社交图谱700，社交网络系统660可以在计算系数时，对连接特定用户节点702和概念节点704的边706的数量和/或类型进行分析。作为示例而非限制，与通过朋友型边连接的用户节点702相比，通过配偶型边(表示两个用户已结婚)连接的用户节点702可以被分配更高的系数。换句话说，取决于分配给特定用户的动作和关系的权重，可以确定关于用户配偶的内容的总亲密度高于关于用户朋友的内容的总亲密度。在特定示例中，用户与另一对象的关系可以影响用户动作的关于计算该对象的系数的权重和/或等级。作为示例而非限制，如果用户在第一照片中被标记，但是该用户仅喜欢第二照片，则社交网络系统660可以确定该用户相对于第一照片的系数高于用户相对于第二照片的系数，因为与内容具有标记类型关系可以被分配比与内容具有喜欢类型关系更高的权重和/或等级。在特定示例中，社交网络系统660可以基于一个或多个第二用户与特定对象的关系，来计算第一用户的系数。换句话说，其他用户与对象的联系和系数可能影响第一用户与该对象的系数。作为示例而非限制，如果第一用户连接到一个或多个第二用户、或者对于一个或多个第二用户具有高系数，并且这些第二用户连接到特定对象、或者对于特定对象具有高系数，则社交网络系统660可以确定第一用户对于该特定对象也应该具有相对高的系数。在特定示例中，该系数可以基于特定对象之间的分离度。较低的系数可以表示第一用户对在社交图谱700中间接连接到该第一用户的用户的内容对象感兴趣的可能性降低。作为示例而非限制，在社交图谱700中更近的社交图谱实体(即，更小的分离度)可以比在社交图谱700中更远的实体具有更高的系数。

[0129] 在特定示例中，社交网络系统660可以基于位置信息来计算系数。在地理上更靠近彼此的对象可以被认为相比于更远的对象而更加彼此相关或对彼此更加感兴趣。在特定示例中，用户对于特定对象的系数可以基于该对象的位置与关联于该用户的当前位置(或者用户的客户端系统630的位置)的接近度。第一用户可能对更接近该第一用户的其他用户或概念更感兴趣。作为示例而非限制，如果用户距离机场一英里并且距离加油站两英里，则社交网络系统660可以基于机场与用户的接近度，来确定用户对于机场比加油站具有更高的系数。

[0130] 在特定示例中,社交网络系统660可以基于系数信息执行关于用户的特定动作。系数可以用于基于用户对特定动作的兴趣,来预测用户是否将执行该特定动作。当生成或向用户呈现任何类型的对象(例如,广告、搜索结果、新闻报道、媒体、消息、通知、或其它合适的对象)时,可以使用系数。在适当的情况下,该系数也可以用于对这些对象进行排名和排序。以这种方式,社交网络系统660可以提供与用户的兴趣和当前环境相关的信息,从而增加他们将找到这种感兴趣的信息的可能性。在特定示例中,社交网络系统660可以基于系数信息生成内容。可以基于特定于用户的系数来提供或选择内容对象。作为示例而非限制,该系数可以用于生成针对用户的媒体,其中,可以向用户呈现该用户相对于媒体对象具有较高的总系数的媒体。作为另一示例而非限制,该系数可以用于生成针对用户的广告,其中,可以向用户呈现该用户相对于广告对象具有较高的总系数的广告。在特定示例中,社交网络系统660可以基于系数信息生成搜索结果。可以基于与关于特定用户的搜索结果相关联的系数,来对该查询用户的搜索结果进行评分或排名。作为示例而非限制,在搜索结果页面上,与具有较高系数的对象相对应的搜索结果的排名可以高于与具有较低系数的对象对应的结果的排名。

[0131] 在特定示例中,社交网络系统660可以响应于对来自特定系统或过程的对系数的请求来计算系数。为了预测用户在给定情况下可能采取的可能动作(或者可能是该动作的主题),任何过程都可以请求针对用户所计算的系数。该请求还可以包括用于计算系数的各种因素所使用的一组权重。该请求可以来自在在线社交网络上运行的过程、来自第三方系统670(例如,通过API或其它通信信道)、或者来自另一合适的系统。响应于该请求,社交网络系统660可以计算系数(或者如果系数信息先前已经被计算和存储,则访问该系数信息)。在特定示例中,社交网络系统660可以测量关于特定过程的亲密度。不同的过程(在线社交网络的内部和外部的过程这两者)可以请求特定对象或一组对象的系数。社交网络系统660可以提供与请求亲密度测量的特定过程相关的亲密度测量结果。以这种方式,每个过程接收针对不同背景定制的亲密度测量结果,在该背景下,过程将使用该亲密度测量结果。

[0132] 与社交图谱亲密度以及亲密度系数相结合,特定示例可以利用于2006年8月11日提交的第11/503093号美国专利申请、于2010年12月22日提交的第12/977027号美国专利申请、于2010年12月23日提交的第12/978265号美国专利申请、以及于2012年10月1日提交的第13/632869号美国专利申请中公开的一个或多个系统、部件、元件、功能、方法、操作或步骤,上述专利申请中的每个专利申请均通过引用被并入。

[0133] 隐私

[0134] 在特定示例中,在线社交网络的多个内容对象中的一个或多个内容对象可以与隐私设置相关联。可以以任何合适的方式存储对象的隐私设置(或“访问设置”),这些方式例如为与该对象相关联、以授权服务器上的索引的方式、以另一合适的方式、或以上方式的任意组合。对象的隐私设置可以指定可以如何使用在线社交网络来访问(例如,查看或共享)该对象(或与该对象相关联的特定信息)。在对象的隐私设置允许特定用户访问该对象的情况下,该对象可以被描述为相对于该用户是“可见的”。作为示例而非限制,在线社交网络的用户可以为用户简档页面指定隐私设置,该隐私设置标识可以访问用户简档页面上的工作经验信息的一组用户,从而拒绝其他用户访问该信息。在特定示例中,隐私设置可以指定不应被允许访问与对象相关联的某些信息的用户的“黑名单”。换句话说,黑名单可以指定这

样的一个或多个用户或实体:对于这些用户或实体,对象是不可见的。作为示例而非限制,用户可以指定不可访问与该用户相关联的相册的一组用户,从而拒绝这些用户访问该相册(同时也可能允许不在该组用户内的某些用户访问该相册)。在特定示例中,隐私设置可以与特定的社交图谱元素相关联。社交图谱元素(例如,节点或边)的隐私设置可以指定可以如何使用在线社交网络来访问该社交图谱元素、与该社交图谱元素相关联的信息、或与该社交图谱元素相关联的内容对象。作为示例而非限制,与特定照片相对于的特定概念节点704可以具有如下隐私设置:该隐私设置指定了该照片可以由该照片中所标记的用户及其朋友访问。在特定示例中,隐私设置可以允许用户选择让或不让其动作被社交网络系统660记录或与其它系统(例如,第三方系统670)共享。在特定示例中,与对象相关联的隐私设置可以指定允许访问或拒绝访问的任何合适的粒度。作为示例而非限制,可以为以下用户指定访问或拒绝访问:特定用户(例如,我、我的室友和我的领导)、特定分离度内的用户(例如,朋友、或朋友的朋友)、用户群组(例如,游戏俱乐部、我的家庭)、用户网(例如,特定雇主的雇员、特定大学的学生或校友)、所有用户(“公共”)、无用户(“私人”)、第三方系统670的用户、特定应用程序(例如,第三方应用程序、外部网站)、其他合适的用户或实体、或以上用户或实体的任意组合。尽管本公开描述了以特定的方式使用特定的隐私设置,但是本公开考虑了以任何合适的方式使用任何合适的隐私设置。

[0135] 在特定示例中,一个或多个服务器662可以是用于实施隐私设置的授权/隐私服务器。响应于来自用户(或其他实体)对存储在数据存储区664中的特定对象的请求,社交网络系统660可以向数据存储区664发送对该对象的请求。该请求可以标识与该请求相关联的用户,并且如果授权服务器基于与该对象相关联的隐私设置确定该用户被授权访问该对象,则该请求可以被发送给该用户(或用户的客户端系统630)。如果请求用户未被授权访问对象,则授权服务器可以阻止从数据存储区664检索所请求的对象,或者可以阻止将所请求的对象发送给该用户。在搜索查询上下文中,如果查询用户被授权访问该对象,则该对象可以作为搜索结果生成。换句话说,该对象可以具有对查询用户可见的可见性。如果对象具有对用户不可见的可见性,则可以从搜索结果中排除该对象。尽管本公开描述了以特定方式执行隐私设置,但本公开考虑了以任何合适的方式执行隐私设置。

[0136] 系统和方法

[0137] 图14示出了示例计算机系统800。系统800可以实现如下项的一个或多个方面:计算架构100(图1)、方法450(图2)、方法470(图3)、方法490(图4)、方法510(图5)、经解压缩的缓冲区200(图6A)、存储器布局250(图6B)、经压缩的块的存储器布局252(图6B)、过程300(图7)、示意图330(图8)、布局350(图9A)、布局370(图9B)、布局390(图9C)、解压缩架构420(图10)和/或高级ANS解压缩流水线510(图11)。在特定示例中,一个或多个计算机系统800执行本文所描述或示出的一种或多种方法中的一个或多个步骤。在特定示例中,一个或多个计算机系统800提供本文所描述或示出的功能。在特定示例中,在一个或多个计算机系统800上运行的软件执行本文所描述或示出的一种或多种方法中的一个或多个步骤,或提供本文所描述或示出的功能。特定示例包括一个或多个计算机系统800的一个或多个部分。在本文中,在适当的情况下,对计算机系统的引用可以包括计算设备,反之亦然。此外,在适当的情况下,对计算机系统的引述可以包括一个或多个计算机系统。

[0138] 本公开考虑了任何合适数量的计算机系统800。本公开考虑了采用任何合适的物

理形式的计算机系统800。作为示例而非限制,计算机系统800可以是:嵌入式计算机系统、片上系统(system-on-chip,SOC)、单板计算机系统(single-board computer system,SBC)(例如,模块上计算机(computer-on-module,COM)或模块上系统(system-on-module,SOM))、台式计算机系统、膝上型或笔记本计算机系统、交互式自助服务终端(kiosk)、大型机、计算机系统网、移动电话、个人数字助理(PDA)、服务器、平板计算机系统、增强/虚拟现实设备、或这些系统中的两者或更多者的组合。在适当的情况下,计算机系统800可以包括一个或多个计算机系统800;可以是单一的或分布式的;可以跨越多个位置;可以跨越多个机器;可以跨越多个数据中心;或可以位于云中,该云可以包括一个或多个网络中的一个或多个云组件。在适当的情况下,一个或多个计算机系统800可以在无实质性的空间限制或时间限制的情况下,执行本文所描述或所示出的一种或多种方法中的一个或多个步骤。作为示例而非限制,一个或多个计算机系统800可以实时地或以分批处理模式执行本文所描述或所示出的一种或多种方法中的一个或多个步骤。在适当的情况下,一个或多个计算机系统800可以在不同的时间或在不同的位置执行本文所描述或所示出的一种或多种方法中的一个或多个步骤。。

[0139] 在特定示例中,计算机系统800包括处理器802、存储器(memory)804、存储装置(storage)806、输入/输出(input/output,I/O)接口808、通信接口810和总线812。尽管本公开描述和示出了具有处于特定布置的特定数量的特定部件的特定计算机系统,但是本公开考虑了具有处于任何合适的布置的任何合适数量的任何合适部件的任何合适的计算机系统。

[0140] 在特定示例中,处理器802包括用于执行多个指令的硬件,这些指令例如为,构成计算机程序的那些指令。作为示例而非限制,为了执行多个指令,处理器802可以从内部寄存器、内部高速缓存、存储器804或存储装置806检索(或读取)这些指令;解码并执行这些指令;然后将一个或多个结果写入内部寄存器、内部高速缓存、存储器804或存储装置806。在特定示例中,处理器802可以包括用于数据、指令或地址的一个或多个内部高速缓存。在适当的情况下,本公开考虑了包括任何合适数量的任何合适的内部高速缓存的处理器802。作为示例而非限制,处理器802可以包括一个或多个指令高速缓存、一个或多个数据高速缓存、和一个或多个转译后备缓冲区(translation lookaside buffer,TLB)。指令高速缓存中的指令可以是存储器804或存储装置806中的指令的副本,并且指令高速缓存可以加速处理器802对这些指令的检索。数据高速缓存中的数据可以是:存储器804或存储装置806中的数据的副本,以供在处理器802处执行的指令在其上操作;在处理器802处执行的先前指令的结果,以供在处理器802处执行的后续指令访问或以用于写入存储器804或存储装置806;或者其它合适的数据。数据高速缓存可以加速由处理器802进行的读取操作或写入操作。TLB可以加速处理器802的虚拟地址转换。在特定示例中,处理器802可以包括用于数据、指令或地址的一个或多个内部寄存器。在适当的情况下,本公开考虑了包括任何合适数量的任何合适的内部寄存器的处理器802。在适当的情况下,处理器802可以包括一个或多个算术逻辑单元(arithmetic logic unit,ALU);可以是多核处理器;或包括一个或多个处理器802。尽管本公开描述和示出了特定的处理器,但是本公开考虑了任何合适的处理器。

[0141] 在特定示例中,存储器804包括主存储器,该主存储器用于存储供处理器802执行的指令或供处理器802操作的数据。作为示例而非限制,计算机系统800可以将多个指令从

存储装置806或另一个源(例如,另一个计算机系统800)加载到存储器804。然后,处理器802可以将这些指令从存储器804加载到内部寄存器或内部高速缓存。为了执行这些指令,处理器802可以从内部寄存器或内部高速缓存检索这些指令并对它们进行解码。在执行指令期间或之后,处理器802可以将一个或多个结果(可以是中间结果或最终结果)写入内部寄存器或内部高速缓存。然后,处理器802可以将这些结果中的一个或多个结果写入存储器804。在特定示例中,处理器802执行一个或多个内部寄存器或一个或多个内部高速缓存中的、或存储器804(与存储装置806不同或其它位置)中的指令,并且对一个或多个内部寄存器或一个或多个内部高速缓存中的、或存储器804(与存储装置806不同或其它位置)中的数据进行操作。一个或多个存储器总线(每条存储器总线可以包括地址总线和数据总线)可以将处理器802耦接到存储器804。如下文所描述的,总线812可以包括一个或多个存储器总线。在特定示例中,一个或多个存储器管理单元(memory management unit,MMU)位于处理器802和存储器804之间,并且促进处理器802所请求的对存储器804的访问。在特定示例中,存储器804包括随机存取存储器(random access memory,RAM)。在适当的情况下,该RAM可以是易失性存储器。在适当的情况下,该RAM可以是动态RAM(dynamic RAM,DRAM)或静态RAM(static RAM,SRAM)。此外,在适当的情况下,该RAM可以是单端口或多端口RAM。本公开考虑了任何合适的RAM。在适当的情况下,存储器804可以包括一个或多个存储器804。尽管本公开描述并示出了特定的存储器,但本公开考虑了任何合适的存储器。

[0142] 在特定示例中,存储装置806包括用于数据或指令的大容量存储装置。作为示例而非限制,存储装置806可以包括硬盘驱动器(hard disk drive,HDD)、软盘驱动器(floppy disk drive,FDD)、闪存、光盘、磁光盘、磁带、或通用串行总线(USB)驱动器、或这些中的两者或更多者的组合。在适当的情况下,存储装置806可以包括可移除介质或不可移除(或固定)介质。在适当的情况下,存储装置806可以处于计算机系统800的内部或外部。在特定示例中,存储装置806是非易失性固态存储器。在特定示例中,存储装置806包括只读存储器(read-only memory,ROM)。在适当的情况下,该ROM可以是掩模编程ROM、可编程ROM(programmable ROM,PROM)、可擦除PROM(erasable PROM,EPROM)、电可擦除PROM(electrically erasable PROM,EEPROM)、电可改写ROM(electrically alterable ROM,EAROM)或闪存,或这些中的两者或更多者的组合。本公开考虑了采用任何合适的物理形式的大容量的存储装置806。在适当的情况下,存储装置806可以包括促进处理器802与存储装置806之间的通信的一个或多个存储控制单元。在适当的情况下,存储装置806可以包括一个或多个存储装置806。尽管本公开描述并示出了特定存储装置,但本公开考虑了任何合适的存储装置。

[0143] 在特定示例中,I/O接口808包括硬件、软件、或硬件和软件这两者,该硬件、软件、或硬件和软件这两者提供用于计算机系统800与一个或多个I/O设备之间的通信的一个或多个接口。在适当的情况下,计算机系统800可以包括这些I/O设备中的一个或多个I/O设备。这些I/O设备中的一个或多个I/O设备可以实现个人与计算机系统800之间的通信。作为示例而非限制,I/O设备可以包括键盘、小键盘、传声器、监视器、鼠标、打印机、扫描仪、扬声器、静物摄像头、手写笔、平板电脑、触摸屏、追踪球、摄像机、另一合适的I/O设备、或这些I/O设备中的两者或更多者的组合。I/O设备可以包括一个或多个传感器。本公开考虑了任何合适的I/O设备和用于这些I/O设备的任何合适的I/O接口808。在适当的情况下,I/O接口

808可以包括如下的一个或多个设备或软件驱动器:所述一个或多个设备或软件驱动器使处理器802能够驱动这些I/O设备中的一个或多个I/O设备。在适当的情况下,I/O接口808可以包括一个或多个I/O接口808。尽管本公开描述和示出了特定的I/O接口,但本公开考虑了任何合适的I/O接口。

[0144] 在特定示例中,通信接口810包括硬件、软件、或硬件和软件这两者,该硬件、软件、或硬件和软件这两者提供用于计算机系统800与一个或多个其它计算机系统800或与一个或多个网络之间的通信(例如,基于数据包的通信)的一个或多个接口。作为示例而非限制,通信接口810可以包括用于与以太网或其它基于有线的网络进行通信的网络接口控制器(network interface controller,NIC)或网络适配器,或者可以包括用于与无线网络(例如,WI-FI网络)进行通信的无线NIC(WNIC)或无线适配器。本公开考虑了任何合适的网络和用于任何合适的网络的任何合适的通信接口810。作为示例而非限制,计算机系统800可以与以下网络进行通信:自组网、个域网(personal area network,PAN)、局域网(LAN)、广域网(WAN)、城域网(MAN)、或互联网的一个或多个部分、或这些中的两个或更多个的组合。这些网络中的一个或多个网络的一个或多个部分可以是有线或无线的。作为示例,计算机系统800可以与无线PAN(wireless PAN,WPAN)(例如,蓝牙(BLUETOOTH)WPAN)、WI-FI网络、WI-MAX网络、蜂窝电话网络(例如,全球移动通信系统(Global System for Mobile Communications,GSM)网络)、或其它合适的无线网络、或这些网络中的两者或更多者的组合进行通信。在适当的情况下,计算机系统800可以包括用于这些网络中的任一者的任何合适的通信接口810。在适当情况下,通信接口810可以包括一个或多个通信接口810。尽管本公开描述和示出了特定的通信接口,但是本公开考虑了任何合适的通信接口。

[0145] 在特定示例中,总线812包括硬件、软件、或硬件和软件这两者,该硬件、软件或硬件和软件这两者将计算机系统800的多个部件彼此耦接。作为示例而非限制,总线812可以包括:加速图形端口(Accelerated Graphics Port,AGP)或其它图形总线、增强型工业标准架构(Enhanced Industry Standard Architecture,EISA)总线、前端总线(front-side bus,FSB)、超传输(HYPERTRANSPORT,HT)互连、工业标准架构(Industry Standard Architecture,ISA)总线、无限带宽(INFINIBAND)互连、低引脚数(low-pin-count,LPC)总线、存储器总线、微通道架构(Micro Channel Architecture,MCA)总线、外围部件互连(Peripheral Component Interconnect,PCI)总线、快速PCI(PCI-Express,PCIe)总线、串行高级技术附件(serial advanced technology attachment,SATA)总线、本地视频电子标准协会(Video Electronics Standards Association local VLB)总线、或另一合适的总线、或这些中的两者或更多者的组合。在适当的情况下,总线812可以包括一个或多个总线812。尽管本公开描述和示出了特定的总线,但本公开考虑了任何合适的总线或互连。

[0146] 示例

[0147] 示例1包括至少一个计算机可读存储介质,该至少一个计算机可读存储介质包括指令集,该指令集在被计算设备执行时使该计算设备:基于第一压缩方案对第一数据进行压缩以生成第二数据,其中,第一数据与第一机器学习模型相关联;将第二数据存储到存储器中;根据第一数据是基于第一压缩方案而被压缩,来调整查找表的第一条目以对应于第一压缩方案;在第一机器学习模型的执行期间,向处理阵列的处理元件提供来自存储器的第二数据;以及在处理阵列处,基于查找表对第二数据进行解压缩以获得第一数据。

[0148] 示例2包括根据示例1所述的至少一个计算机可读存储介质,其中,该指令集在被执行时使该计算设备:基于第二压缩方案对第三数据进行压缩以生成第四数据,其中,第三数据与第二机器学习模型相关联;将第四数据存储到存储器中;根据第三数据是基于第二压缩方案而被压缩,来调整查找表的第二条目以对应于第二压缩方案;在第二机器学习模型的执行期间,向处理阵列提供来自存储器的第三数据;以及在处理阵列处,基于查找表对第四数据进行解压缩以获得第三数据。

[0149] 示例3包括根据示例2所述的至少一个计算机可读存储介质,其中,该指令集在被执行时使计算设备:用第二条目替换查找表中的第一条目。

[0150] 示例4包括根据示例2所述的至少一个计算机可读存储介质,其中,该指令集在被执行时使该计算设备:将第二条目插入到查找表中,使得查找表包括第一条目和第二条目。

[0151] 示例5包括根据示例1所述的至少一个计算机可读存储介质,其中,该指令集在被执行时使该计算设备:基于第一压缩方案,识别第一数据的第一部分能够压缩一量;以及基于该量满足压缩阈值,确定要用第一压缩方案对第一部分进行压缩。

[0152] 示例6包括根据示例1所述的至少一个计算机可读存储介质,其中,该指令集在被执行时使该计算设备:确定用第一压缩方案对第一数据进行压缩将第一数据的大小减小到小于或等于存储器的可用存储器空间;预测第一压缩方案被应用至少第一次数;确定该至少第一次数满足重复使用阈值;基于该至少第一次数满足重复使用阈值,和用第一压缩方案对第一数据进行压缩将第一数据的大小减小到小于或等于可用存储器空间,来选择第一压缩方案;生成元数据以将第一压缩方案与第二数据相关联;以及将第二数据与元数据相关联地存储到存储器中。

[0153] 示例7包括根据示例1至6中任一项所述的至少一个计算机可读存储介质,其中,第一数据包括权重或偏置中的一者或多者;第一压缩方案是非对称数字系统压缩方案;第一数据是在第一机器模型被执行之前基于第一压缩方案来压缩的;以及为了对第二数据进行解压缩,处理阵列用于使用处理元件对第二数据进行解压缩。

[0154] 示例8包括一种系统,该系统包括:一个或多个处理器;以及第一存储器,该第一存储器耦接到该一个或多个处理器,第一存储器包括由该一个或多个处理器执行的指令,该一个或多个处理器在执行这些指令时能够操作以:基于第一压缩方案对第一数据进行压缩以生成第二数据,其中,第一数据与第一机器学习模型相关联;将第二数据存储到第二存储器中;根据第一数据是基于第一压缩方案而被压缩,来调整查找表的第一条目以对应于第一压缩方案;在执行第一机器学习模型期间,向处理阵列的处理元件提供来自第二存储器的第二数据;以及在处理阵列处,基于查找表对第二数据进行解压缩以获得第一数据。

[0155] 示例9包括根据示例8所述的系统,其中,该一个或多个处理器在执行这些指令时,还能够操作以:基于第二压缩方案对第三数据进行压缩以生成第四数据,其中,第三数据与第二机器学习模型相关联;将第四数据存储到第二存储器中;根据第三数据是基于第二压缩方案而被压缩,来调整查找表的第二条目以对应于第二压缩方案;在第二机器学习模型的执行期间,向处理阵列提供来自第二存储器的第三数据;以及在处理阵列处,基于查找表对第四数据进行解压缩以获得第三数据。

[0156] 示例10包括根据示例9所述的系统,其中,该一个或多个处理器在执行这些指令时还能够操作以:用第二条目替换查找表中的第一条目。

[0157] 示例11包括根据示例9所述的系统,其中,该一个或多个处理器在执行指令时还能够操作以:将第二条目插入到查找表中,使得查找表包括第一条目和第二条目。

[0158] 示例12包括根据示例8所述的系统,其中,该一个或多个处理器在执行这些指令时还能够操作以:基于第一压缩方案,识别第一数据的第一部分能够压缩一量;以及基于该量满足压缩阈值,确定要用第一压缩方案对第一部分进行压缩。

[0159] 示例13包括根据示例8所述的系统,其中,该一个或多个处理器在执行这些指令时还能够操作以:确定用第一压缩方案对第一数据进行压缩将第一数据的大小减小到小于或等于第二存储器的可用存储器空间;预测第一压缩方案被应用至少第一次数;确定该至少第一次数满足重复使用阈值;基于该至少第一次数满足重复使用阈值,和用第一压缩方案对第一数据进行压缩将第一数据的大小减小到小于或等于可用存储器空间,来选择第一压缩方案;生成元数据以将第一压缩方案与第二数据相关联;以及将第二数据与元数据相关联地存储到第二存储器中。

[0160] 示例14包括根据示例8至13中任一项所述的系统,其中,第一数据包括权重或偏置中的一者或多者;第一压缩方案是非对称数字系统压缩方案;第一数据是在第一机器模型被执行之前基于第一压缩方案来压缩的;并且为了对第二数据进行解压缩,处理阵列用于使用处理元件对第二数据进行解压缩。

[0161] 示例15包括一种方法,该方法包括:基于第一压缩方案对第一数据进行压缩以生成第二数据,其中,第一数据与第一机器学习模型相关联;将第二数据存储到存储器中;根据第一数据是基于第一压缩方案而被压缩,来调整查找表的第一条目以对应于第一压缩方案;在第一机器学习模型的执行期间,向处理阵列的处理元件提供来自存储器的第二数据;以及在处理阵列处,基于查找表对第二数据进行解压缩以获得第一数据。

[0162] 示例16包括根据示例15所述的方法,该方法还包括:基于第二压缩方案对第三数据进行压缩以生成第四数据,其中,第三数据与第二机器学习模型相关联;将第四数据存储到存储器中;根据第三数据是基于第二压缩方案而被压缩,来调整查找表的第二条目以对应于第二压缩方案;在第二机器学习模型的执行期间,向处理阵列提供来自存储器的第三数据;以及在处理阵列处,基于查找表对第四数据进行解压缩以获得第三数据。

[0163] 示例17包括根据示例16所述的方法,该方法还包括:用第二条目替换查找表中的第一条目。

[0164] 示例18包括根据示例16所述的方法,该方法还包括:将第二条目插入到查找表中,使得查找表包括第一条目和第二条目。

[0165] 示例19包括根据示例15所述的方法,该方法还包括:基于第一压缩方案,识别第一数据的第一部分能够压缩一量;以及基于该量满足压缩阈值,确定要用第一压缩方案对第一部分进行压缩。

[0166] 示例20包括根据示例15至19中任一项所述的方法,该方法还包括:确定用第一压缩方案对第一数据进行压缩将第一数据的大小减小到小于或等于存储器的可用存储器空间;预测第一压缩方案被应用至少第一次数;确定该至少第一次数满足重复使用阈值;基于该至少第一次数满足重复使用阈值,和用第一压缩方案对第一数据进行压缩将第一数据的大小减小到小于或等于可用存储器空间,来选择第一压缩方案;生成元数据以将第一压缩方案与第二数据相关联;以及将与元数据相关联的第二数据存储到存储器中;其中,第一数

据包括权重或偏置中的一者或多者;第一压缩方案是非对称数字系统压缩方案;第一数据是在第一机器模型被执行之前基于第一压缩方案来压缩的;以及为了对第二数据进行解压缩,处理阵列用处理元件对第二数据进行解压缩。

[0167] 示例21包括一种系统,该系统包括用于基于第一压缩方案对第一数据进行压缩以生成第二数据的装置,其中,第一数据与第一机器学习模型相关联;用于将第二数据存储到存储器中的装置;用于根据第一数据是基于第一压缩方案而被压缩来调整查找表的第一条目以对应于第一压缩方案的装置;用于在第一机器学习模型的执行期间向处理阵列的处理元件提供来自存储器的第二数据的装置;以及用于在处理阵列处基于查找表对第二数据进行解压缩以获得第一数据的装置。

[0168] 示例22包括根据示例21所述的系统,该系统还包括:用于基于第二压缩方案对第三数据进行压缩以生成第四数据的装置,其中,第三数据与第二机器学习模型相关联;用于将第四数据存储到存储器中的装置;用于根据第三数据是基于第二压缩方案而被压缩来调整查找表的第二条目以对应于第二压缩方案的装置;用于在第二机器学习模型的执行期间向处理阵列提供来自存储器的第三数据的装置;以及用于在处理阵列处基于查找表对第四数据进行解压缩以获得第三数据的装置。

[0169] 示例23包括根据示例22所述的系统,该系统还包括:用于用第二条目替换查找表中的第一条目的装置。

[0170] 示例24包括根据示例22所述的系统,该系统还包括:用于将第二条目插入到查找表中使得查找表包括第一条目和第二条目的装置。

[0171] 示例25包括根据示例21所述的系统,该系统还包括:用于基于第一压缩方案识别第一数据的第一部分能够压缩一量的装置;以及,用于基于该量满足压缩阈值来确定要用第一压缩方案对第一部分进行压缩的装置。

[0172] 示例26包括根据示例21所述的系统,该系统还包括:用于确定用第一压缩方案对第一数据进行压缩将第一数据的大小减小到小于或等于存储器的可用存储器空间的装置;用于预测第一压缩方案被应用至少第一次数的装置;用于确定该至少第一次数满足重复使用阈值的装置;用于基于该至少第一次数满足重复使用阈值,和用第一压缩方案对第一数据进行压缩将第一数据的大小减小到小于或等于可用存储器空间,来选择第一压缩方案的装置;用于生成元数据以将第一压缩方案与第二数据相关联的装置;以及用于将第二数据与元数据相关联地存储到存储器中的装置。

[0173] 示例27包括根据示例21至26所述的系统中的任何一个,其中,第一数据包括权重或偏置中的一者或多者;第一压缩方案是非对称数字系统压缩方案;第一数据是在第一机器模型被执行之前基于第一压缩方案来压缩的;以及为了对第二数据进行解压缩,处理阵列用于使用处理元件对第二数据进行解压缩。

[0174] 在本文中,在适当的情况下,一个或多个计算机可读非暂态存储介质可以包括:一个或多个基于半导体的集成电路(integrated circuit,IC)或其它集成电路(例如,现场可编程门阵列(field-programmable gate array,FPGA)或专用IC(application-specific IC,ASIC))、硬盘驱动器(HDD)、混合硬盘驱动器(hybrid hard drive,HHD)、光盘、光盘驱动器(optical disc drive,ODD)、磁光盘、磁光驱动器、软盘、软盘驱动器(FDD)、磁带、固态驱动器(solid-state drive,SSD)、RAM驱动器、安全数字卡或驱动器、任何其它合适的计算机

可读非暂态存储介质、或这些中的两者或更多者的任何合适的组合。在适当的情况下,计算机可读非暂态存储介质可以是易失性的、非易失性的、或者易失性和非易失性的组合。

[0175] 示例适用于与所有类型的半导体集成电路(integrated circuit,"IC")芯片一起使用。这些IC芯片的示例包括但不限于处理器、控制器、芯片组部件、可编程逻辑阵列(programmable logic array,PLA)、存储器芯片、网络芯片、片上系统(SOC)、SSD/NAND控制器ASIC等。此外,在一些附图中,信号导线用线表示。一些可以不同以指示更多组成信号路径,具有号码标签以指示多个组成信号路径,和/或在一个或多个端具有箭头以指示主要信息流方向。然而,这不应以限制的方式来解释。相反,此类添加的细节可以连同一个或多个示例性示例来使用以促进对电路的更简单理解。任何表示的信号线,无论是否具有附加信息,都可以实际上包括可在多个方向中行进和可用任何适合类型的信号方案实现的一个或更多信号,例如,用差动对、光纤线、和/或单个端的线来实现的数字或模拟线。

[0176] 可能已经给出示例大小/模型/值/范围,但示例不限于此。因为制造技术(例如,光刻法)随着时间而成熟,所以期望的是,可以制造更小尺寸的设备。另外,IC芯片和其它部件的公知的功率/接地连接为了说明和讨论的简单性而可以被或可以不被示出在附图内,并以致不模糊示例的某些方面。另外,布置可以以框图形式被示出,以便避免模糊示例并且还考虑到关于实现此类框图布置的细节高度取决于计算系统(该示例要在其内实现)的事实,即,此类细节应该充分在本领域技术人员的专业范畴内。在特定细节(例如,电路)被阐明以便描述示例性示例的情况下,对本领域技术人员应显而易见的是,示例可以在没有这些特定细节或有这些特定细节的变化的情况下被实践。因此,该描述被认为是说明性的而不是限制性的。

[0177] 术语"耦合/耦接"在本文中可以用来指所讨论的部件之间任何类型的直接或间接的关系,并可以应用于电学、机械、流体、光学、电磁、电机或其它连接。此外,术语"第一"、"第二"等在本文中可以用来促进讨论,并不带有任何具体的时间或时间顺序的显著性意义,除非另外指明。

[0178] 如在本申请和权利要求书中使用的,由术语"中的一个或多个"联接的项的列表可意指所列项的任何组合。例如,短语"A、B或C中的一个或多个"可意指A;B;C;A和B;A和C;B和C;或A、B和C。

[0179] 本领域技术人员将从前述描述中理解,示例的广泛技术可以以各种形式实现。因此,虽然已经结合其多个示例中的特定示例描述了该多个示例,但是该多个示例的真实范围不应如此受限,因为在研究附图、说明书和所附的权利要求书时,对于技术人员来说,其它修改将变得显而易见。

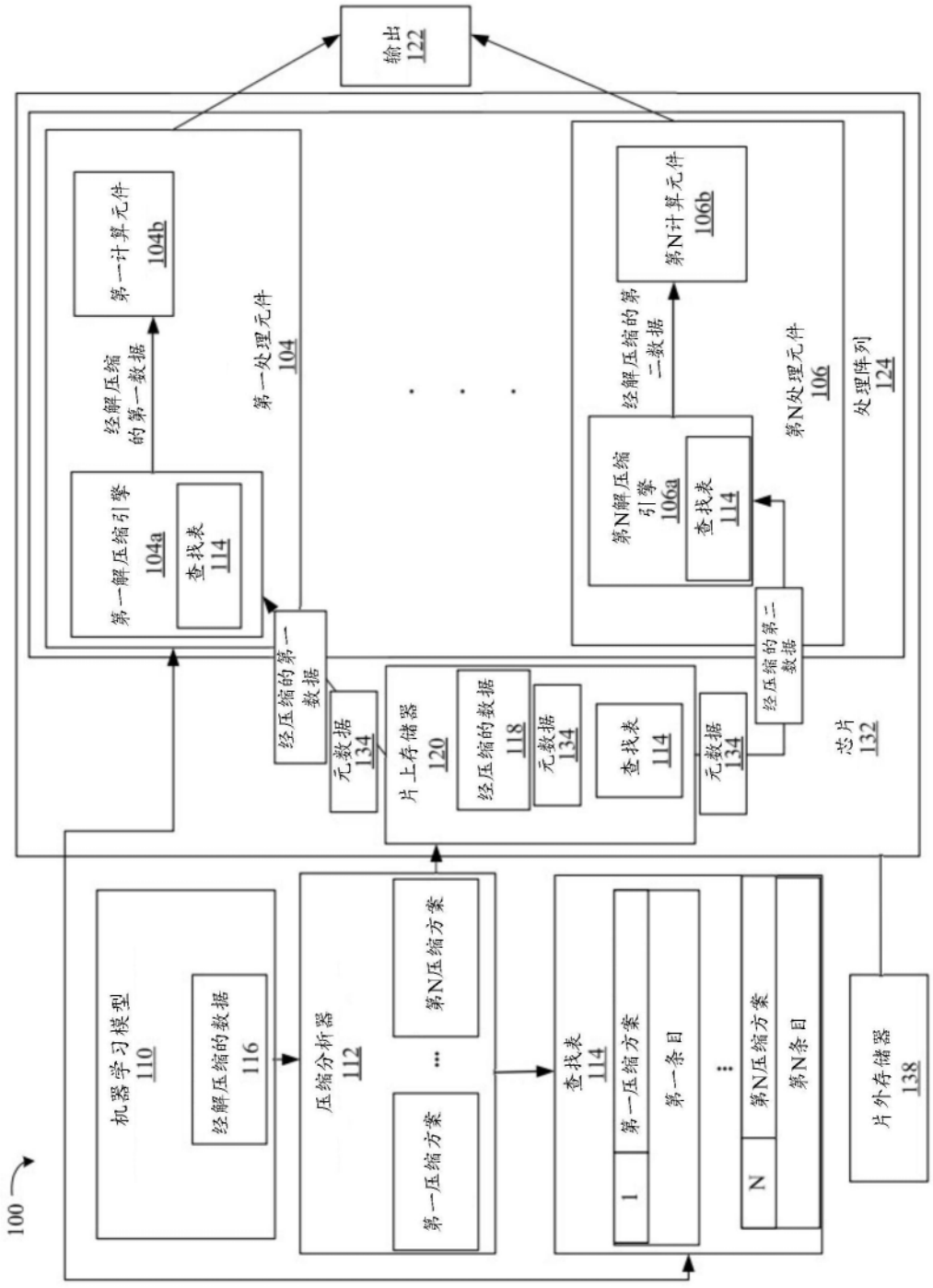

100
机器学习模型
110
经解压缩的数据
116
压缩分析器
112
第一压缩方案
...
第N压缩方案
查找表
114
1
第一压缩方案
第一条目
N
第N压缩方案
第N条目
片外存储器
138
输出
122
第一处理元件
104
第一解压缩引擎
104a
查找表
114
经解压缩
的第一数据
第一计算元件
104b
经压缩的第一
数据
元数据
134
片上存储器
120
经压缩的数据
118
元数据
134
查找表
114
元数据
134
经压缩的第二
数据
芯片
132
第N处理元件
106
第N解压缩
引擎
106a
查找表
114
经解压缩的第
二数据
第N计算元件
106b
处理阵列
124

图1

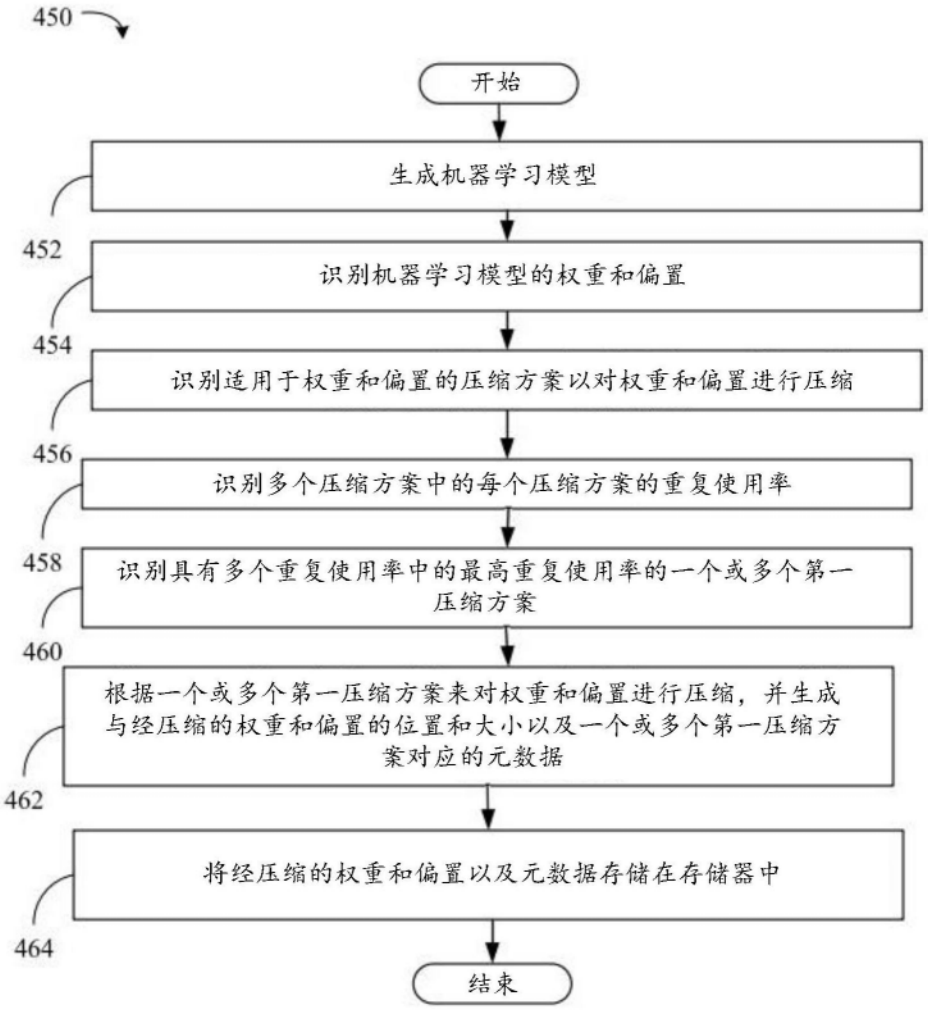

450
开始
生成机器学习模型
452
识别机器学习模型的权重和偏置
454
识别适用于权重和偏置的压缩方案以对权重和偏置进行压缩
456
识别多个压缩方案中的每个压缩方案的重复使用率
458
识别具有多个重复使用率中的最高重复使用率的一个或多个第一压缩方案
460
根据一个或多个第一压缩方案来对权重和偏置进行压缩，并生成与经压缩的权重和偏置的位置和大小以及一个或多个第一压缩方案对应的元数据
462
将经压缩的权重和偏置以及元数据存储在存储器中
464
结束

图2

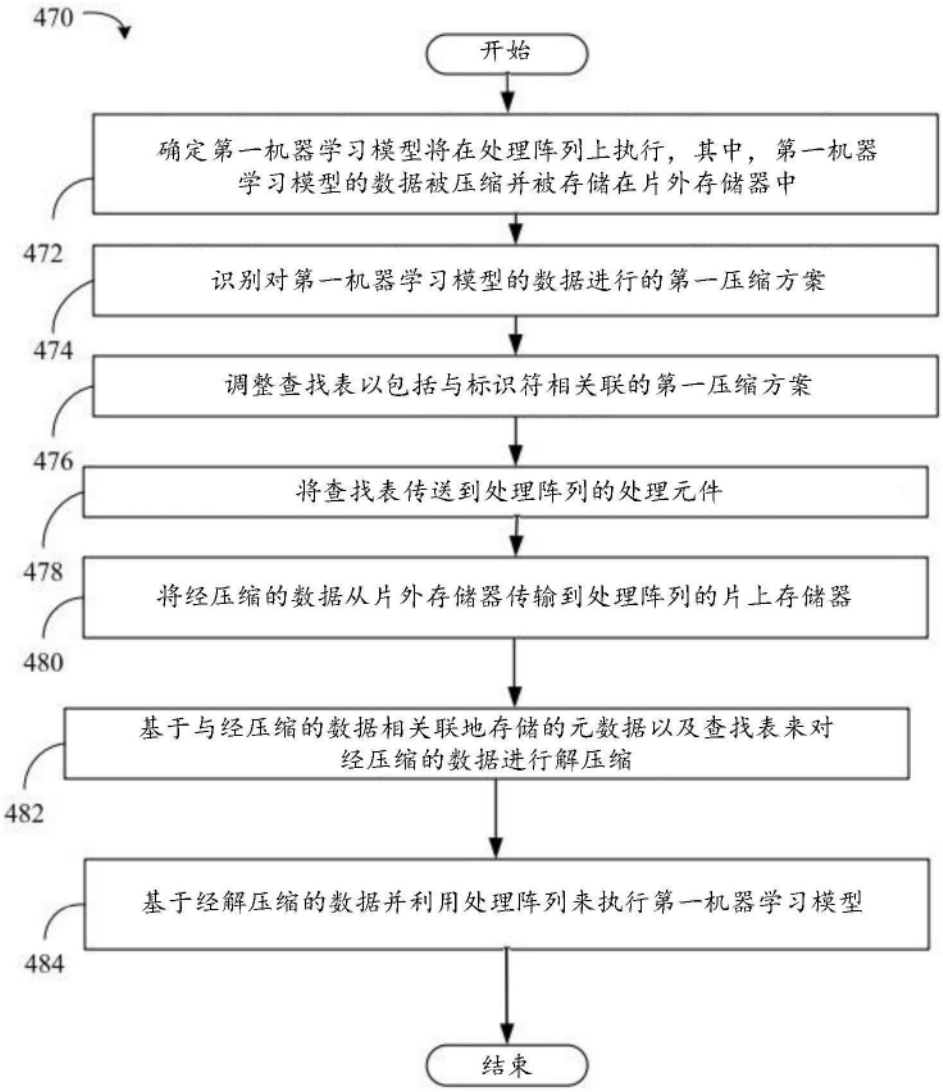

470
开始
确定第一机器学习模型将在处理阵列上执行，其中，第一机器学习模型的数据被压缩并被存储在片外存储器中
472
识别对第一机器学习模型的数据进行的第一压缩方案
474
调整查找表以包括与标识符相关联的第一压缩方案
476
将查找表传送到处理阵列的处理元件
478
将经压缩的数据从片外存储器传输到处理阵列的片上存储器
480
基于与经压缩的数据相关联地存储的元数据以及查找表来对经压缩的数据进行解压缩
482
基于经解压缩的数据并利用处理阵列来执行第一机器学习模型
484
结束

图3

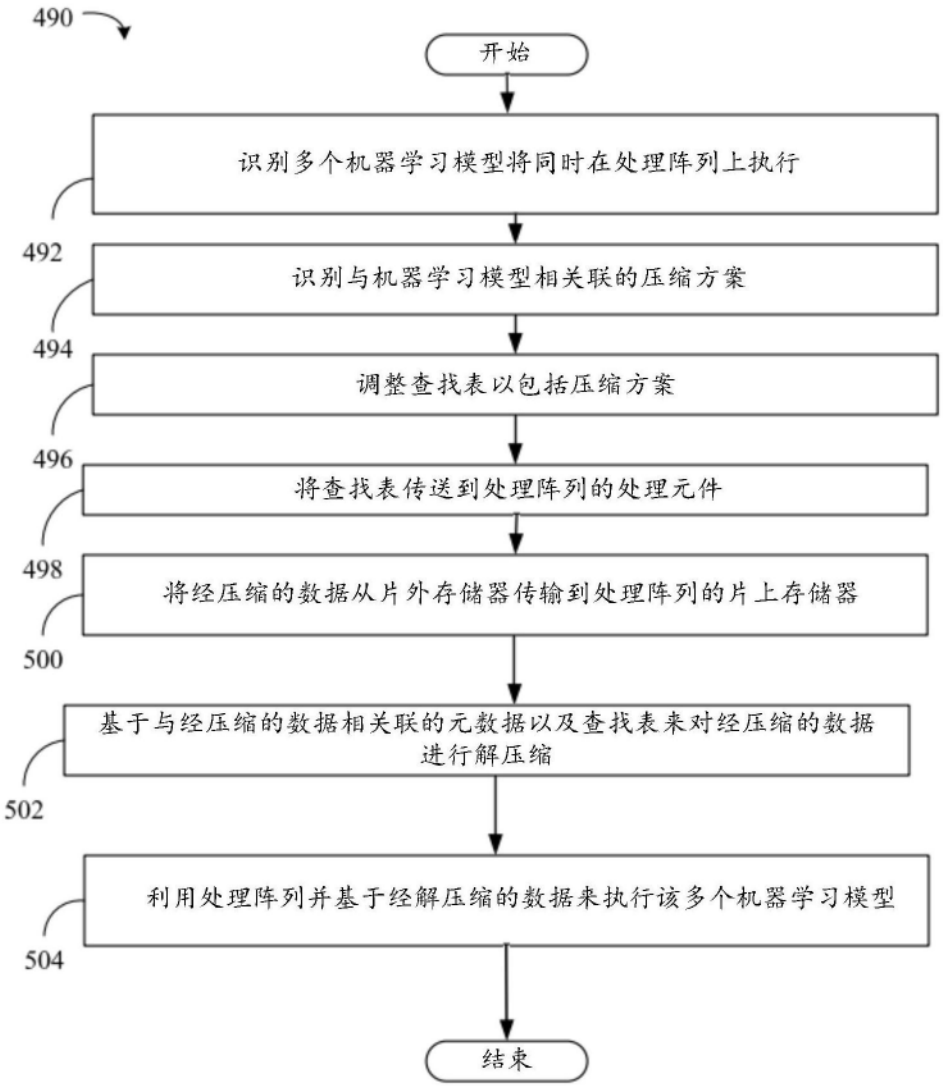

490
开始
识别多个机器学习模型将同时在处理阵列上执行
492
识别与机器学习模型相关联的压缩方案
494
调整查找表以包括压缩方案
496
将查找表传送到处理阵列的处理元件
498
将经压缩的数据从片外存储器传输到处理阵列的片上存储器
500
基于与经压缩的数据相关联的元数据以及查找表来对经压缩的数据进行解压缩
502
利用处理阵列并基于经解压缩的数据来执行该多个机器学习模型
504
结束

图4

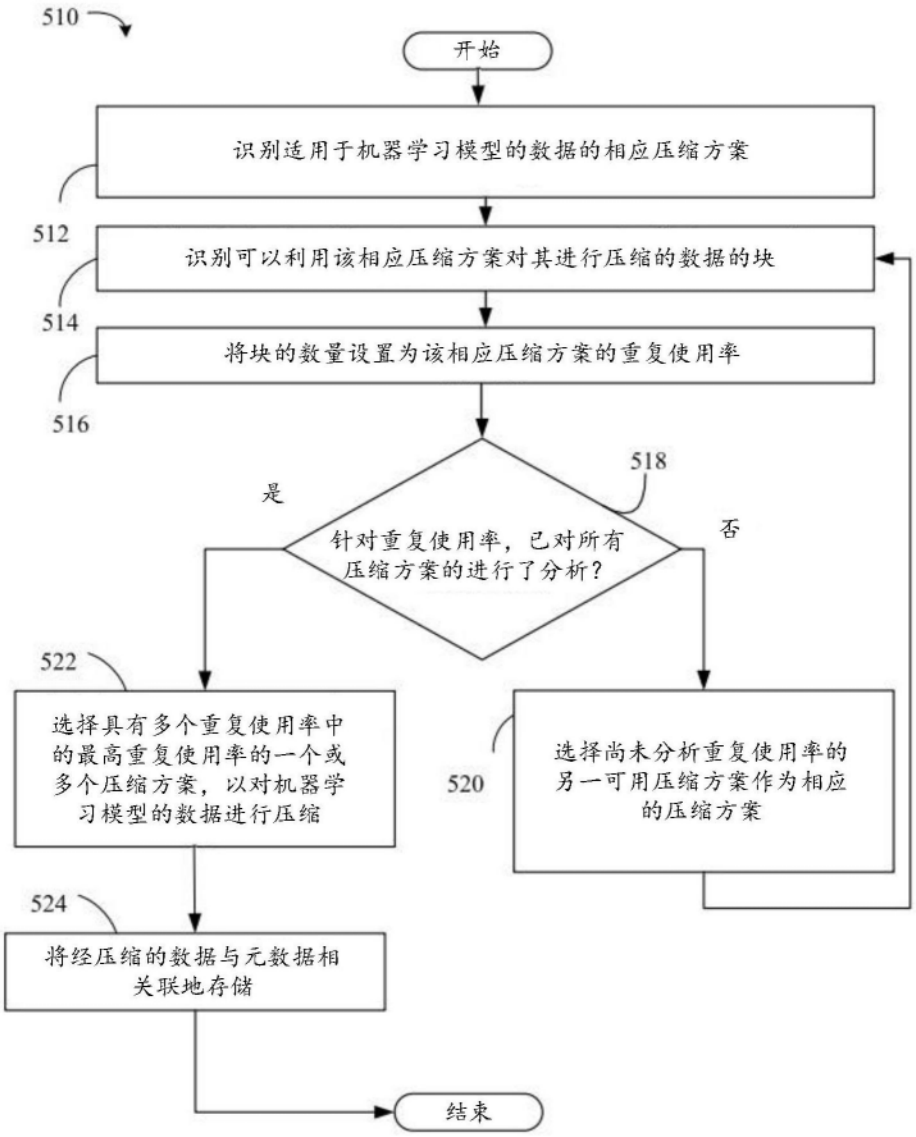

510
开始
识别适用于机器学习模型的数据的相应压缩方案
512
识别可以利用该相应压缩方案对其进行压缩的数据的块
514
将块的数量设置为该相应压缩方案的重复使用率
516
518
是
否
针对重复使用率，已对所有压缩方案的进行了分析？
522
选择具有多个重复使用率中的最高重复使用率的一个或多个压缩方案，以对机器学习模型的数据进行压缩
520
选择尚未分析重复使用率的另一可用压缩方案作为相应的压缩方案
524
将经压缩的数据与元数据相关联地存储
结束

图5

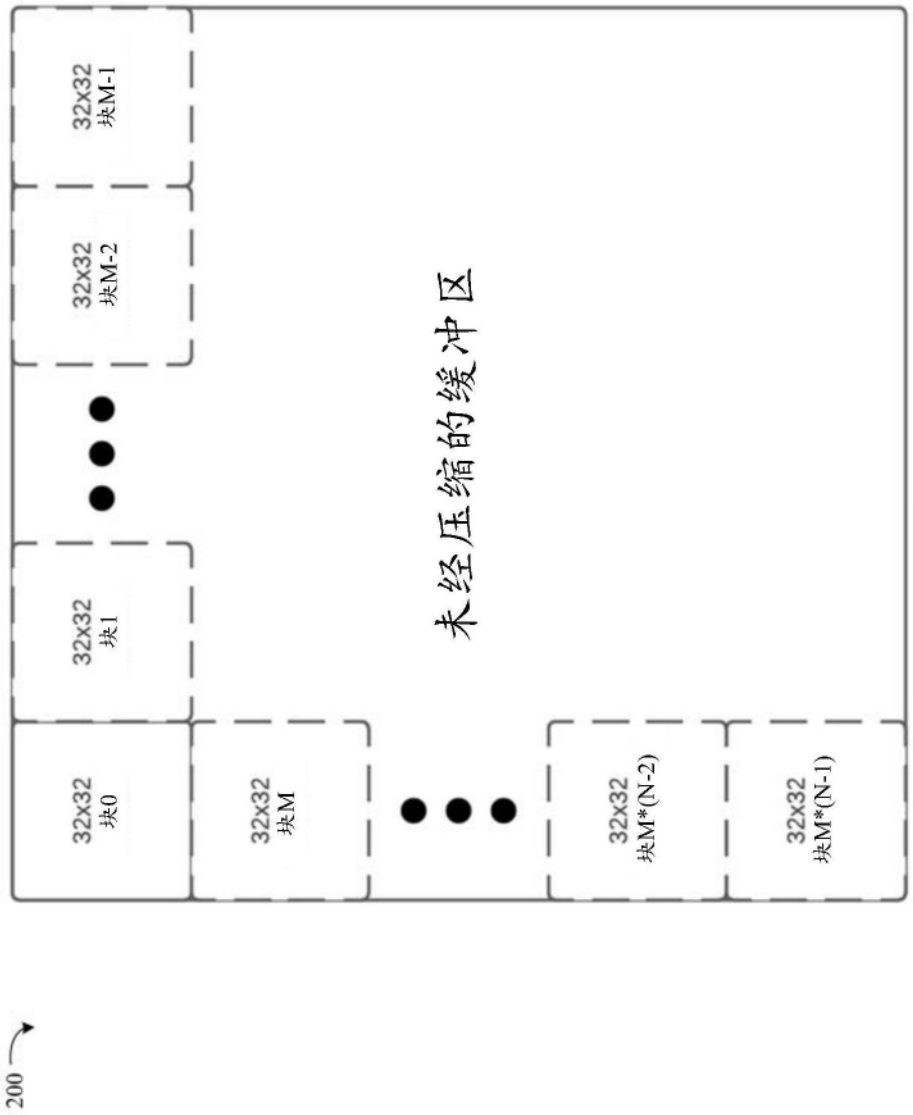

未经压缩的缓冲区
32x32 块0
32x32 块1
32x32 块M
32x32 块M-1
32x32 块M-2
32x32 块M*(N-2)
32x32 块M*(N-1)
200

图6A

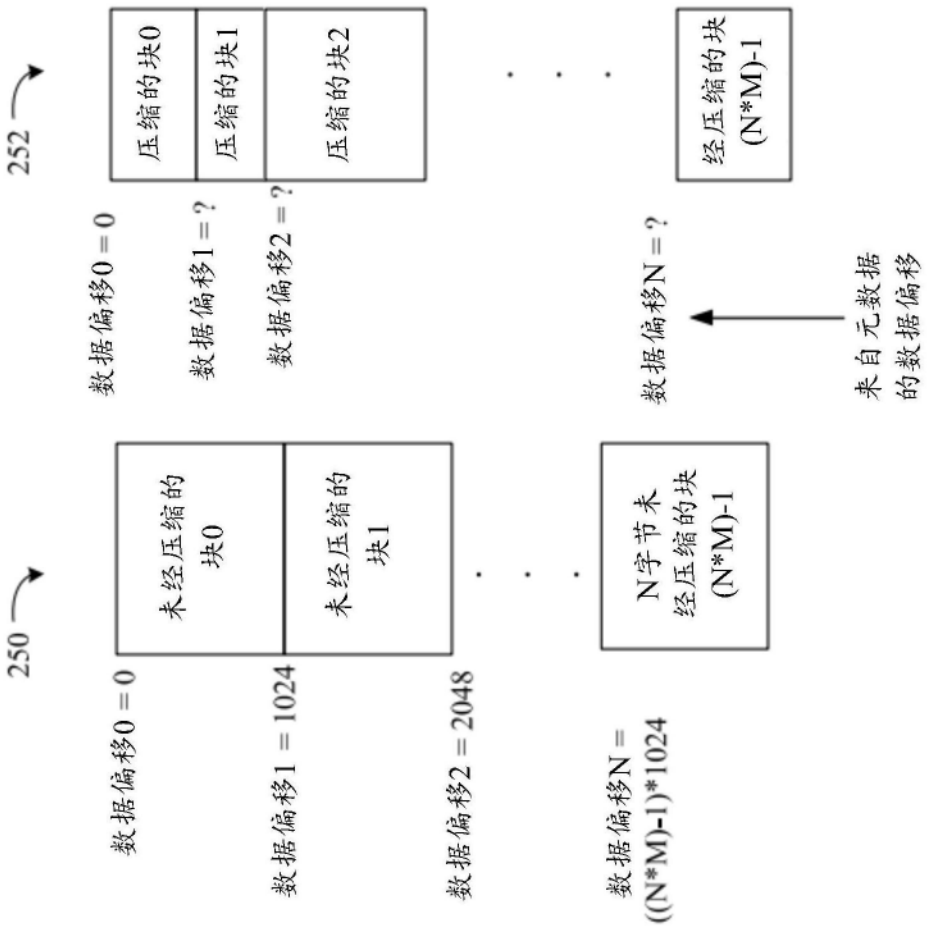

252
压缩的块0
压缩的块1
压缩的块2
经压缩的块
(N*M)-1
数据偏移0 = 0
数据偏移1 = ?
数据偏移2 = ?
数据偏移N = ?
来自元数据
的数据偏移
250
未经压缩的
块0
未经压缩的
块1
N字节未
经压缩的块
(N*M)-1
数据偏移0 = 0
数据偏移1 = 1024
数据偏移2 = 2048
数据偏移N =
((N*M)-1)*1024

图6B

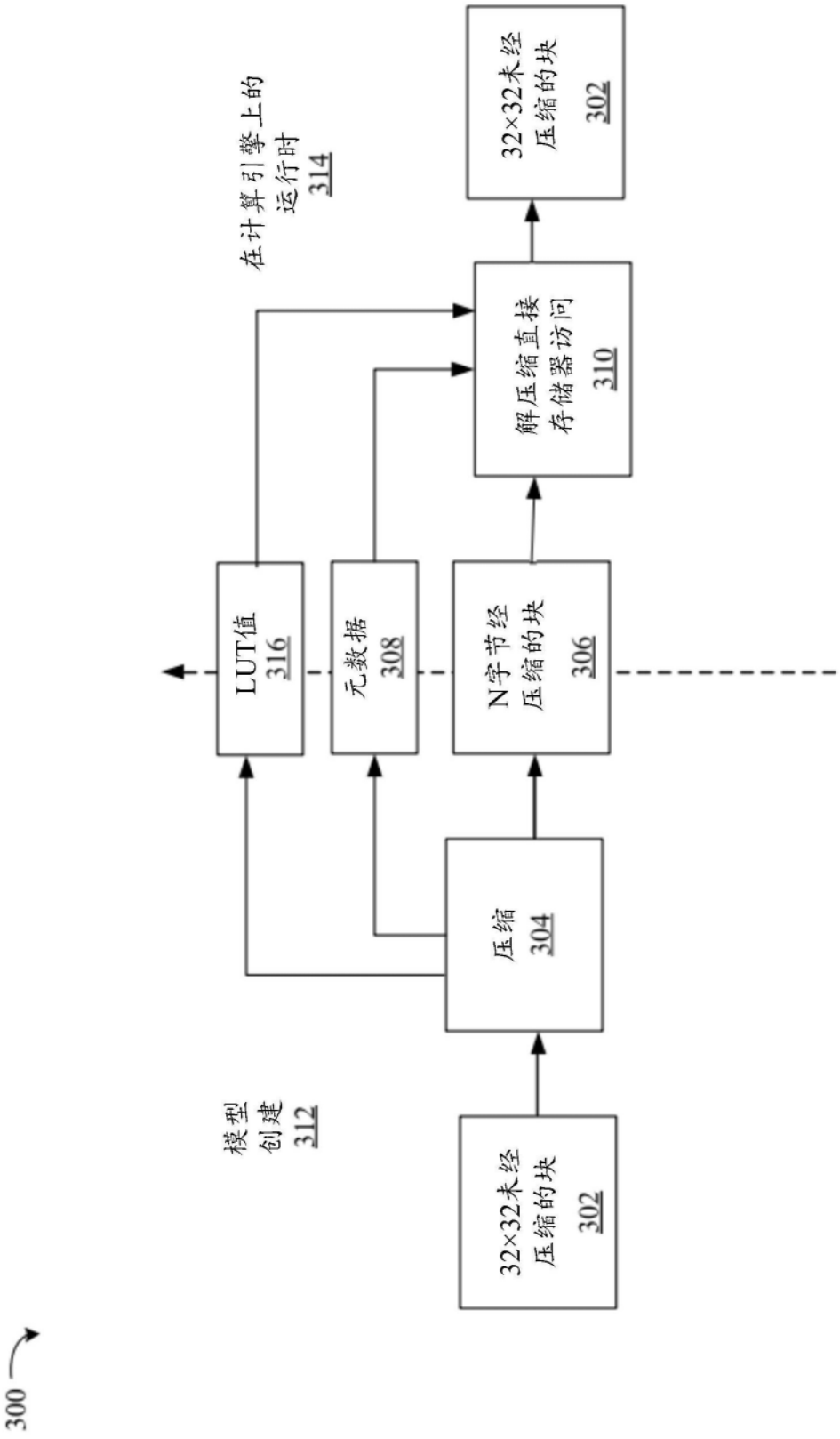

300
32×32未经压缩的块 302
模型创建 312
压缩 304
32×32未经压缩的块 302
在计算引擎上的运行时 314
LUT值 316
元数据 308
N字节经压缩的块 306
解压缩直接存储器访问 310
32×32未经压缩的块 302

图7

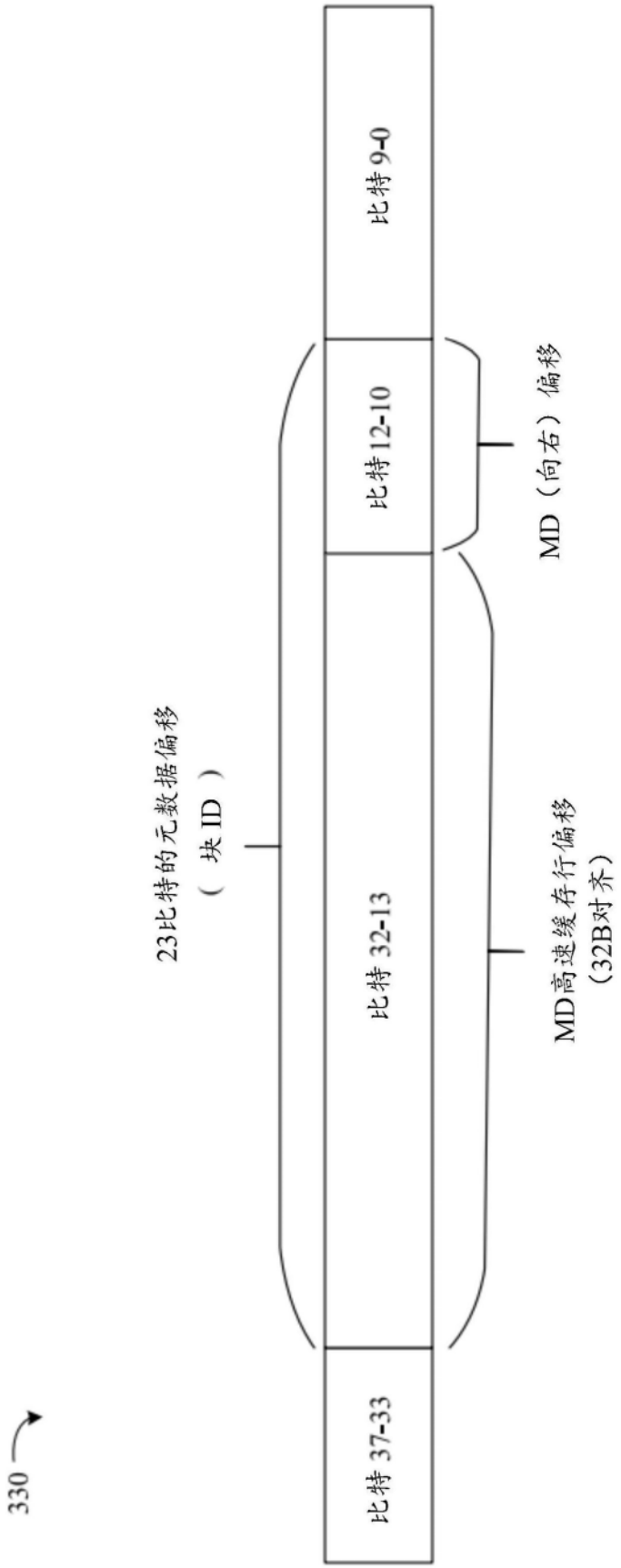

330
比特 37-33
比特 32-13
比特 12-10
比特 9-0
23比特的元数据偏移
（ 块 ID ）
MD高速缓存行偏移
（32B对齐）
MD（向右）偏移

图8

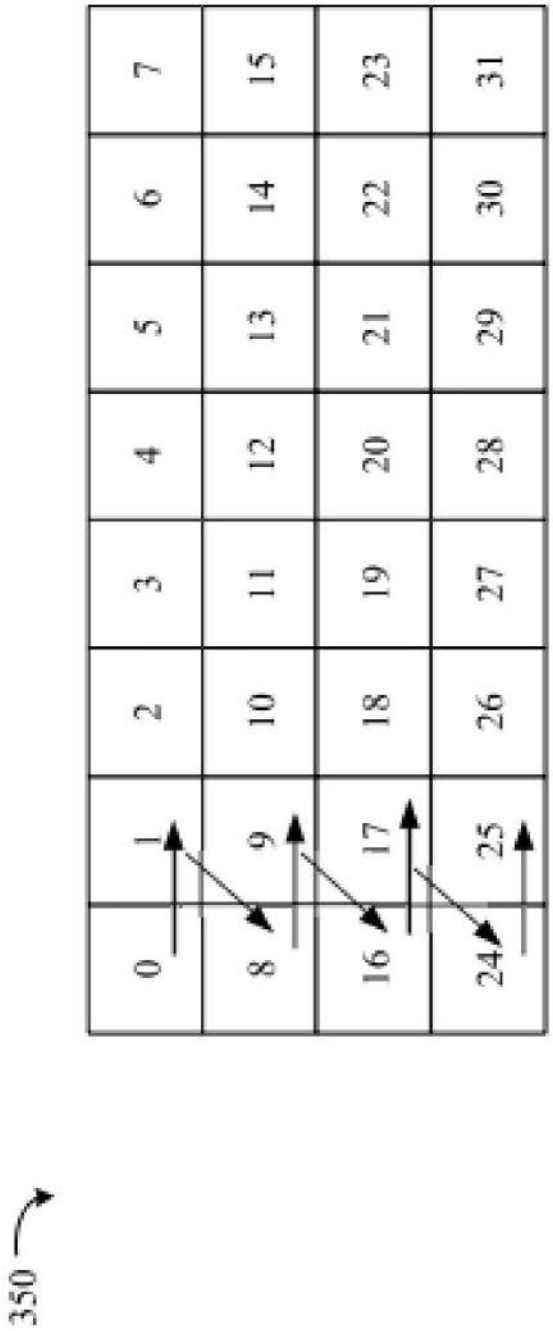

7
15
23
31
6
14
22
30
5
13
21
29
4
12
20
28
3
11
19
27
2
10
18
26
1
9
17
25
0
8
16
24
350

图9A

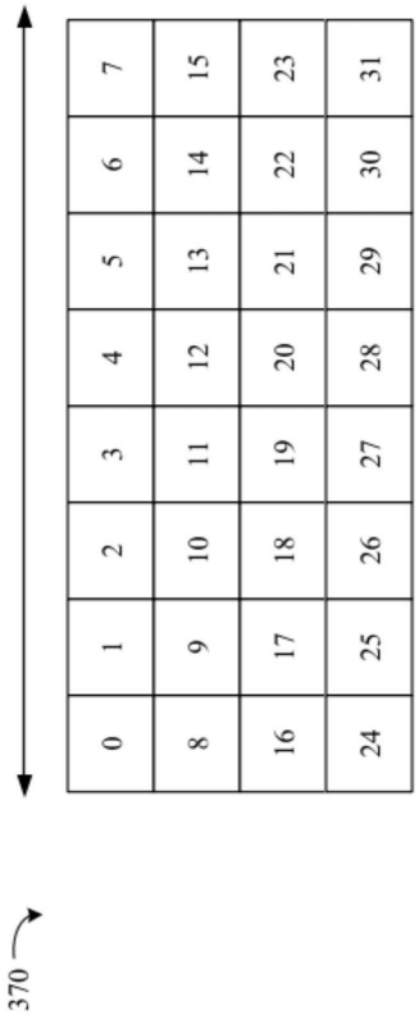

0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
370

图9B

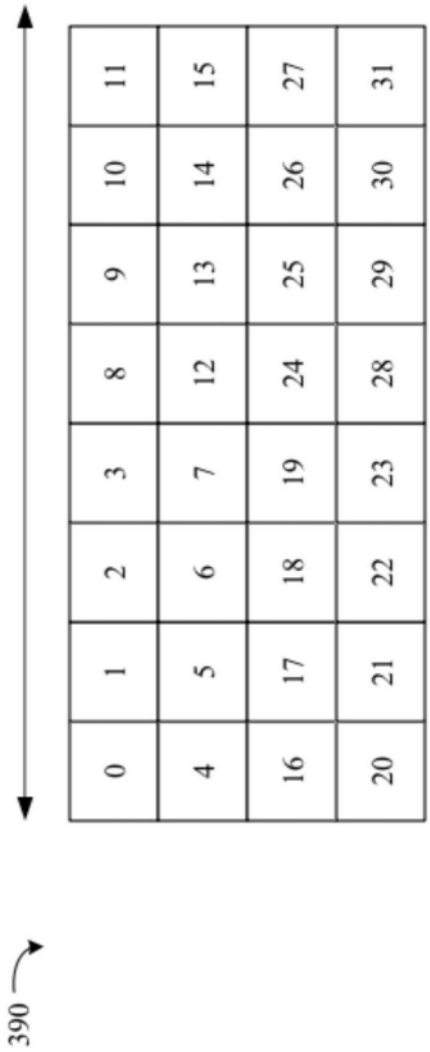

0
1
2
3
8
9
10
11
4
5
6
7
12
13
14
15
16
17
18
19
24
25
26
27
20
21
22
23
28
29
30
31
390

图9C

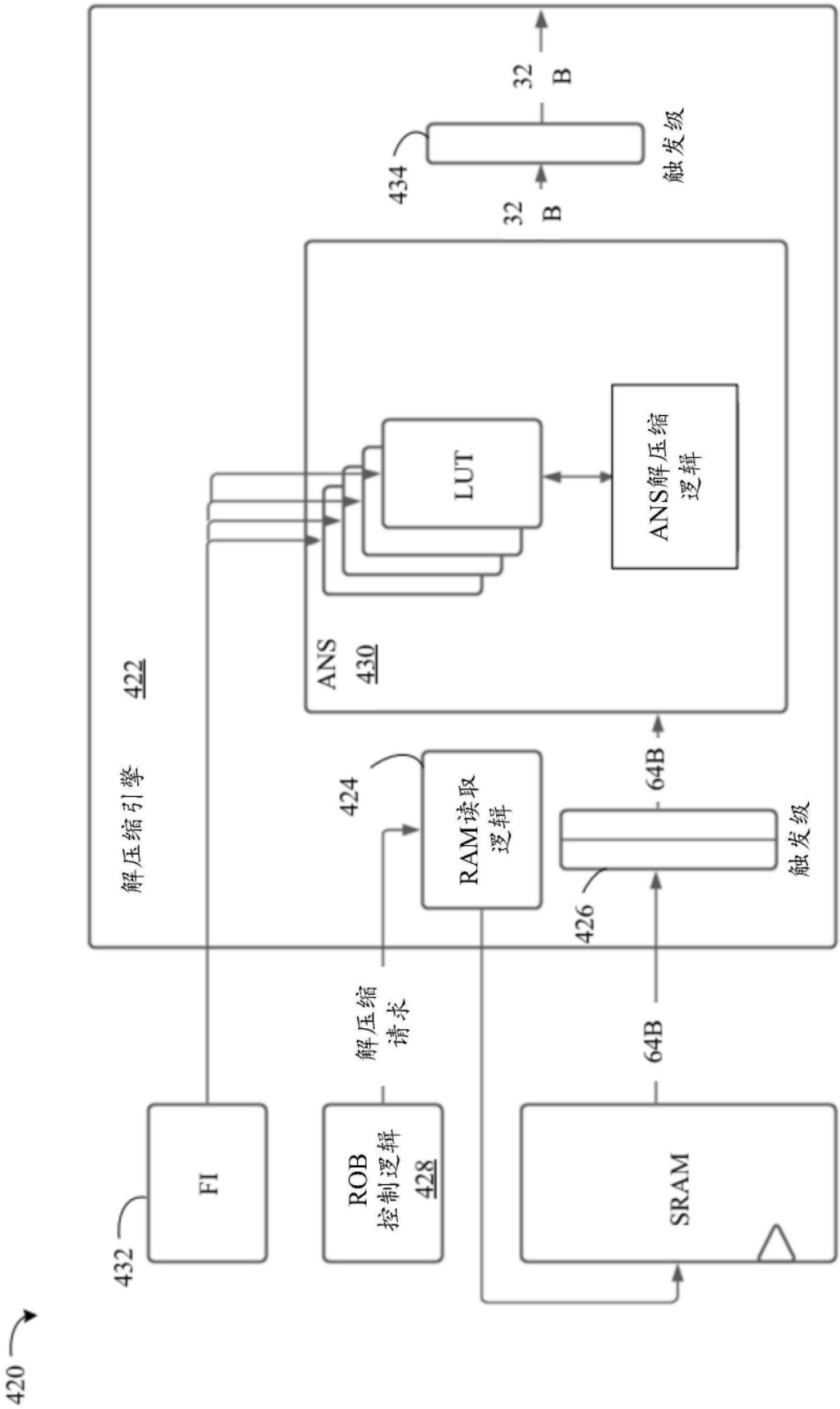

420
解压缩引擎 422
FI
432
ROB 控制逻辑 428
解压缩请求
SRAM
64B
RAM读取逻辑
424
426
触发级
64B
ANS 430
LUT
ANS解压缩逻辑
434
32 B
32 B
触发级

图10

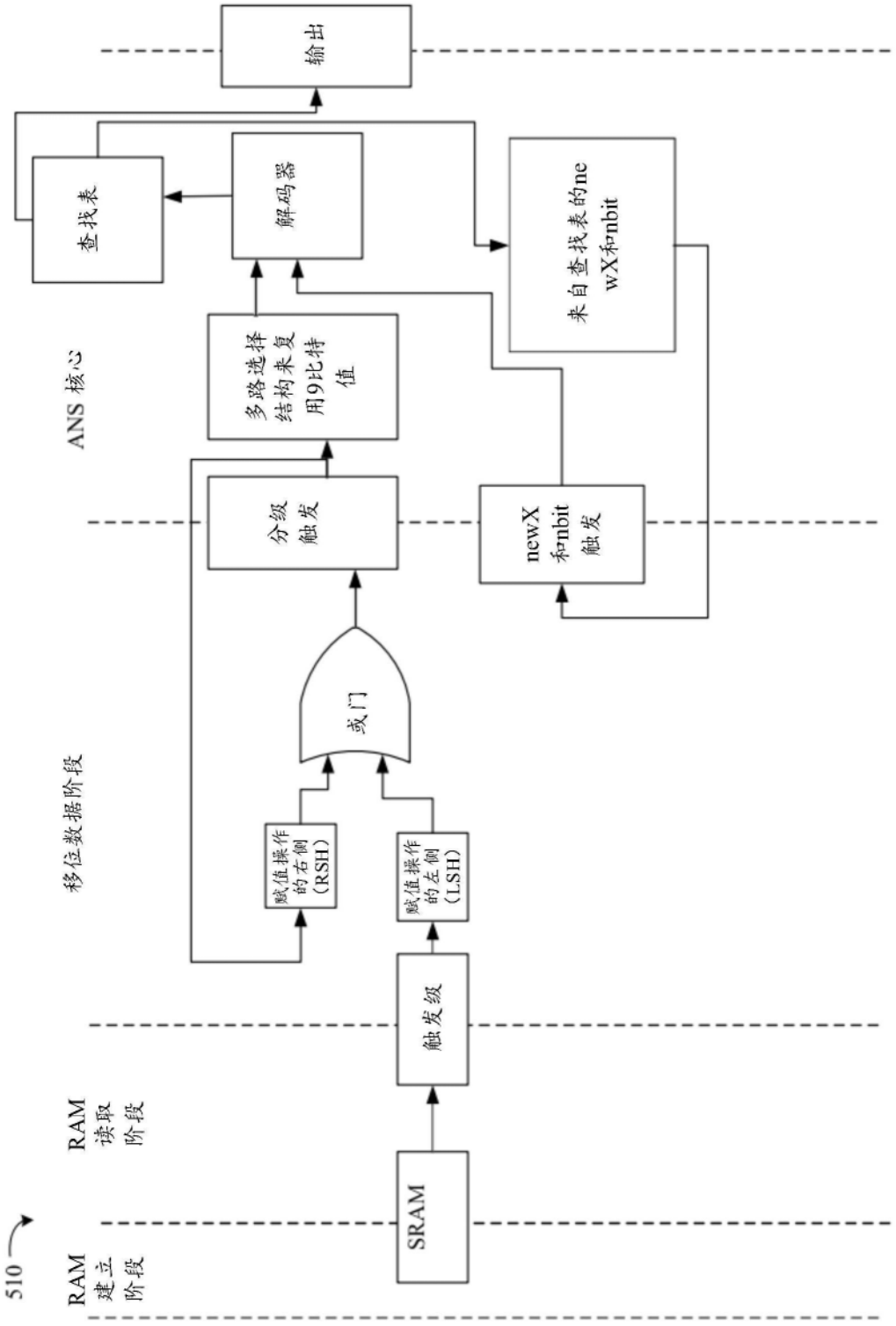

510
RAM 建立阶段
RAM 读取阶段
移位数据阶段
ANS 核心
SRAM
触发级
赋值操作的左侧（LSH）
赋值操作的右侧（RSH）
或门
分级触发
多路选择结构来复用9比特值
newX和nbit触发
来自查找表的newX和nbit
解码器
查找表
输出

图11

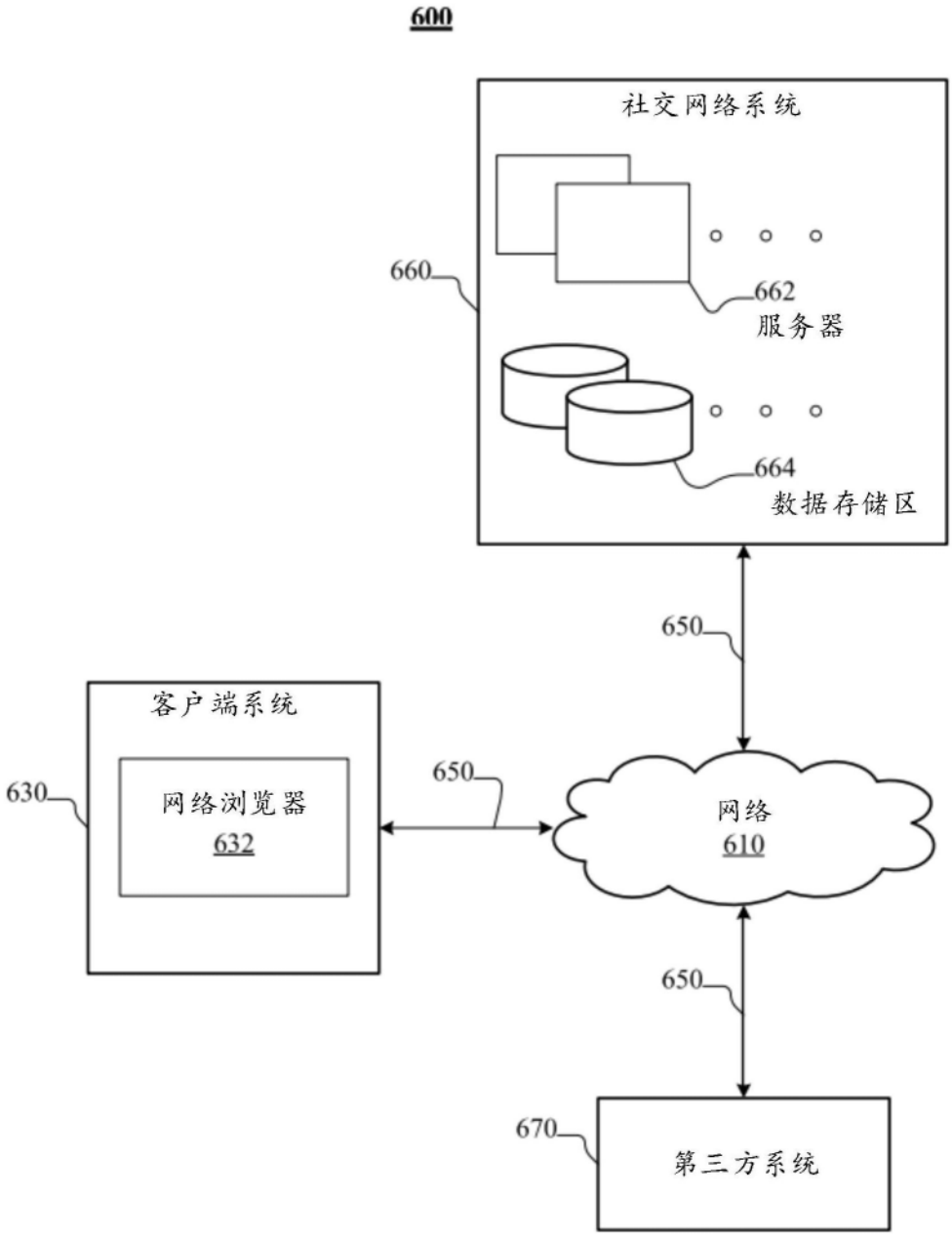

600
社交网络系统
660
662
服务器
664
数据存储区
650
客户端系统
630
网络浏览器
632
650
网络
610
650
670
第三方系统

图12

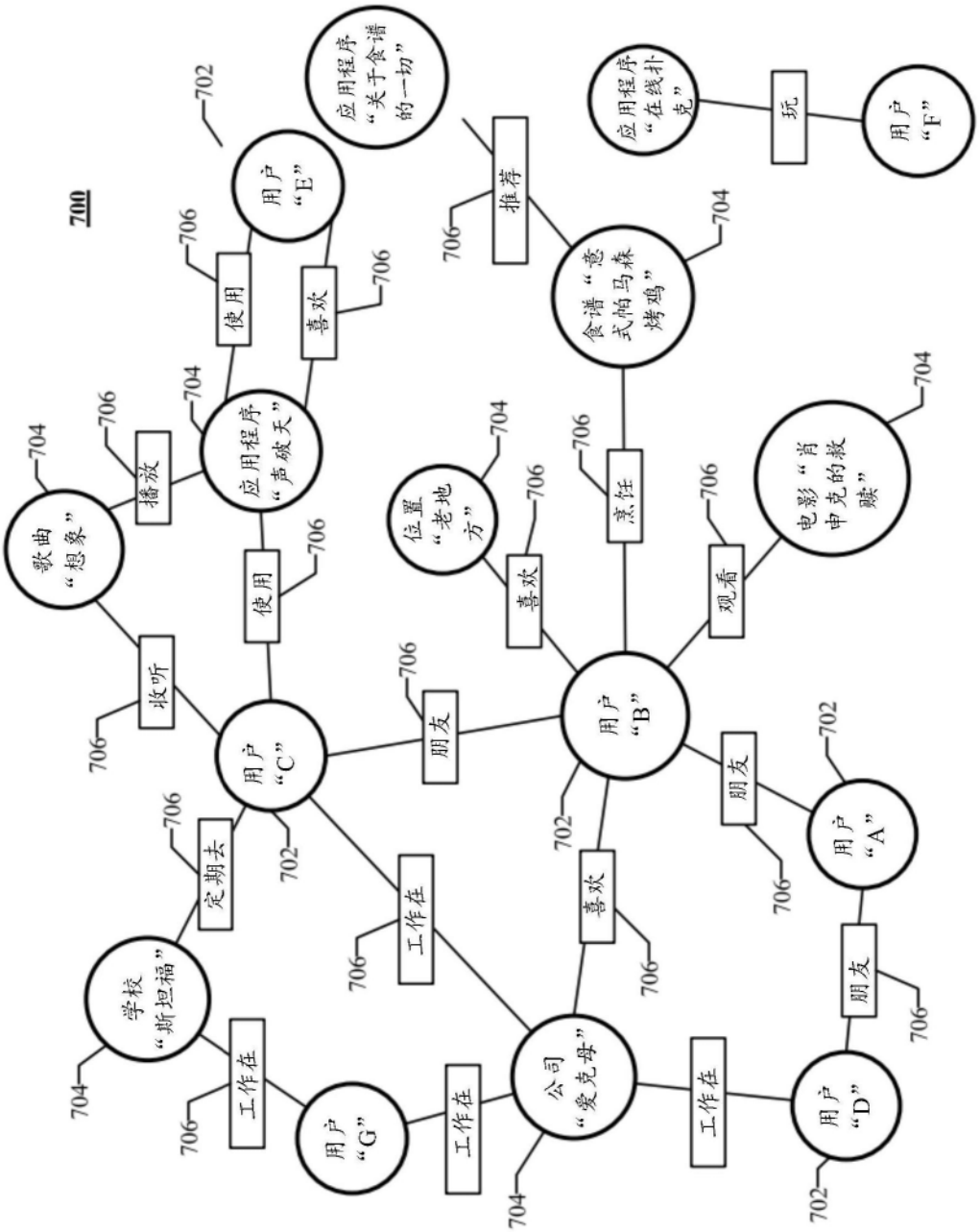

700
用户 "E"
应用程序 "关于食谱的一切"
使用
喜欢
应用程序 "声破天"
播放
歌曲 "想象"
使用
收听
用户 "C"
朋友
用户 "B"
推荐
食谱 "意式帕马森烤鸡"
应用程序 "在线扑克"
玩
用户 "F"
位置 "老地方"
喜欢
烹饪
观看
电影 "肖申克的救赎"
朋友
用户 "A"
定期去
学校 "斯坦福"
工作在
用户 "G"
工作在
公司 "爱克母"
喜欢
工作在
朋友
用户 "D"

图13

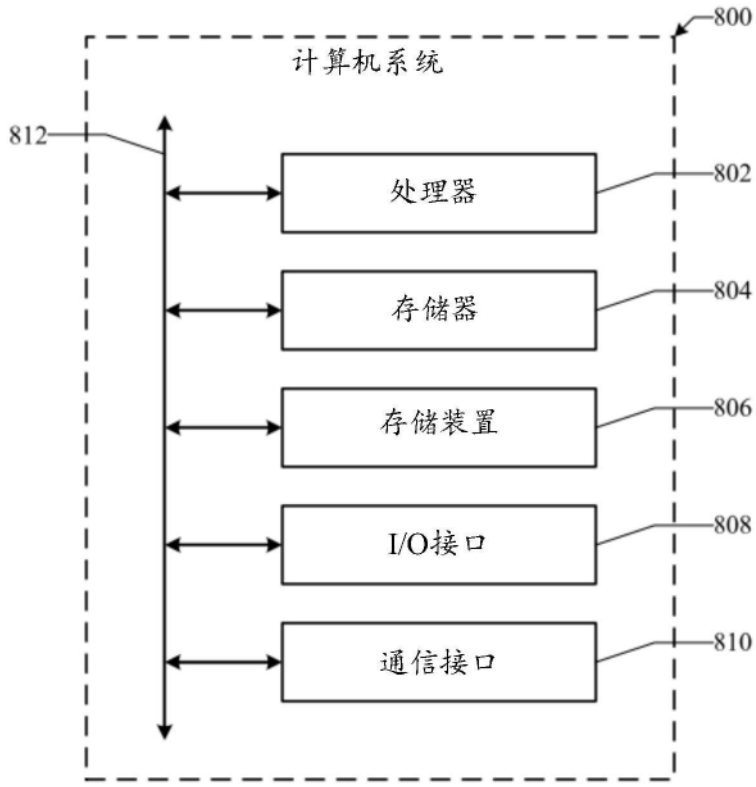


图14