(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 111353591 A
(43)申请公布日 2020.06.30

(21)申请号 201811566331.6

(22)申请日 2018.12.20

(71)申请人 中科寒武纪科技股份有限公司
地址 100000 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 广州三环专利商标代理有限公司 44202
代理人 郝传鑫 熊永强

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

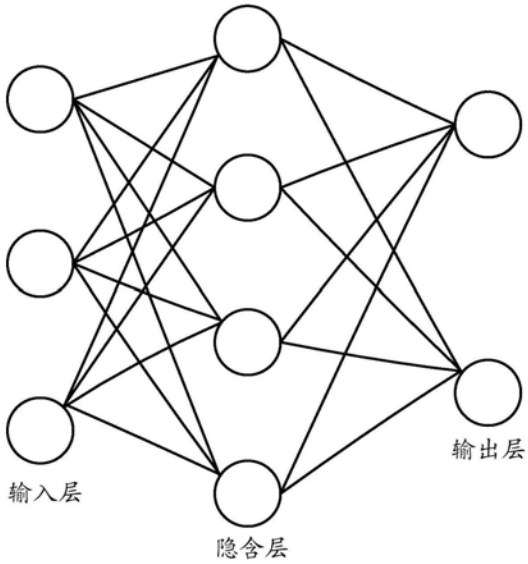
权利要求书5页 说明书25页 附图15页

(54)发明名称

一种计算装置及相关产品

(57)摘要

本申请提供了一种计算装置及相关产品，所述计算装置包括压缩单元、运算单元以及控制器单元；其中，控制器单元，用于获取针对第一输入数据的压缩请求，并根据压缩请求指示压缩单元对第一输入数据进行压缩；其中，第一输入数据包括第一权值矩阵；压缩单元，用于将第一权值矩阵压缩为第二权值矩阵；控制器单元，还用于根据第二输入数据以及计算指令执行神经网络计算。通过本申请，在神经网络压缩过程中，可以保证神经网络模型的拓扑结构保持不变，从而避免了神经网络模型的拓扑结构出现不规则，减少了神经网络的运算量。



CN 111353591 A

1.一种计算装置,其特征在于,所述计算装置用于执行机器学习计算,所述计算装置包括:压缩单元、运算单元以及控制器单元;

所述控制器单元,用于获取针对第一输入数据的压缩请求,并根据所述压缩请求指示所述压缩单元对所述第一输入数据进行压缩;其中,所述第一输入数据包括第一权值矩阵;

所述压缩单元,用于将第一权值矩阵压缩为第二权值矩阵;

所述控制器单元,还用于获取第二输入数据以及计算指令;所述第二输入数据包括所述第二权值矩阵以及输入神经元数据;

所述控制器单元,还用于解析该计算指令得到多个运算指令,将所述多个运算指令以及所述第二输入数据发送给运算单元;

所述运算单元,用于获取所述运算指令,并根据所述运算指令以及所述第二输入数据执行神经网络计算。

2.根据权利要求1所述的计算装置,其特征在于,所述压缩单元包括:分解单元、求解单元和训练单元;

其中,所述分解单元,用于将所述第一权值矩阵分解成第三权值矩阵;其中,所述第三权值矩阵包括至少两个子矩阵;

所述求解单元,用于根据第一公式确定所述至少两个子矩阵中的每个子矩阵的大小,所述第一公式为$Q\approx Q_1*Q_2*.....*Q_n$;其中,所述Q表示第一权值矩阵;所述$Q_1$表示所述至少两个子矩阵中的第一子矩阵;所述$Q_2$表示所述至少两个子矩阵中的第二子矩阵;所述$Q_n$表示所述至少两个子矩阵中的第n子矩阵;

所述训练单元,用于调整所述至少两个子矩阵中的每个子矩阵的大小,并通过训练压缩后的机器学习模型,以得到满足预设精度的第二权值矩阵。

3.根据权利要求2所述的计算装置,其特征在于,所述求解单元,用于根据第一公式确定所述至少两个子矩阵中的每个子矩阵,所述第一公式为$Q\approx Q_1*Q_2*.....*Q_n$,包括:

所述求解单元,具体用于根据所述第一公式和第二公式确定所述至少两个子矩阵中的每个子矩阵的大小,所述第二公式为$||Q-Q_1*Q_2*.....*Q_n||\leqslant T$,其中,所述T表示预设的误差阈值。

4.根据权利要求2所述的计算装置,其特征在于,所述训练单元,用于调整所述至少两个子矩阵中的每个子矩阵的大小,并通过训练压缩后的机器学习模型,以得到满足预设精度的第二权值矩阵,包括:

所述训练单元,具体用于调整所述至少两个子矩阵中的每个子矩阵的大小,并通过训练压缩后的机器学习模型,以得到满足预设精度并且与所述第一权值矩阵之间的压缩比满足预设压缩比的第二权值矩阵。

5.根据权利要求2至4任一项所述的计算装置,其特征在于,所述计算装置用于执行全连接层神经网络计算;所述至少两个子矩阵包括两个子矩阵;所述第一公式包括:$M\approx M_1*M_2$;所述两个子矩阵包括第一子矩阵M1和第二子矩阵M2,所述M1为$N_{in}*K$矩阵,所述M2为$K*N_{out}$矩阵;其中,K为压缩参数,N_{in}为所述神经网络的输入神经元的个数,N_{out}为所述神经网络的输出神经元的个数;所述压缩参数用于表征所述M1的输出神经元的个数以及所述M2的输入神经元的个数,所述K为大于0且小于等于$\min(N_{in},N_{out})$的正整数。

6.根据权利要求2-4任一项所述的计算装置,其特征在于,所述计算装置用于执行卷积

层神经网络计算;所述卷积层神经网络包括$N_{fin}*N_{fout}$个卷积核;所述第一公式包括:$F \approx F_1 * F_2$;其中,F表示所述$N_{fin}*N_{fout}$个卷积核中的任意一个卷积核;所述F1为第一子卷积核;所述F2为第二子卷积核;所述第一子卷积核F1为(K_x,R),所述第二子卷积核F2为(R,K_y),(K_x,K_y)表示卷积核的大小,R为压缩参数,所述R为大于0且小于等于$\min(K_x,K_y)$的正整数。

7.根据权利要求2-4任一项所述的计算装置,其特征在于,所述计算装置用于执行LSTM层神经网络计算,所述LSTM层包括N个全连接层,所述N为大于0的正整数;针对第j个全连接层,所述第一公式包括:$M_j \approx M_{j\_1} * M_{j\_2}$;所述第j个全连接层中的两个子矩阵包括第一子矩阵$M_{j_1}$和第二子矩阵$M_{j_2}$,所述$M_{j_1}$为$N_{in_j}*S$矩阵,所述$M_{j_2}$为$S*N_{out_j}$矩阵;其中,S为压缩参数,$N_{in_j}$为所述神经网络第j个全连接层的输入神经元的个数,$N_{out_j}$为所述神经网络第j个全连接层的输出神经元的个数;所述压缩参数用于表征所述M_{j_1}的输出神经元的个数以及所述M_{j_2}的输入神经元的个数,所述S为大于0且小于等于$\min(N_{in_j},N_{out_j})$的正整数。

8.根据权利要求1所述的计算装置,其特征在于,其特征在于,所述运算单元包括:一个主处理电路和多个从处理电路;

所述主处理电路对所述第二输入数据执行前序处理以及与所述多个从处理电路之间传输数据和运算指令;

所述多个从处理电路根据从所述主处理电路传输的数据以及运算指令并行执行中间运算得到多个中间结果,并将所述多个中间结果传输给所述主处理电路;

所述主处理电路对所述多个中间结果执行后续处理得到所述计算指令的计算结果。

9.根据权利要求1所述的计算装置,其特征在于,所述计算装置还包括:存储单元和直接内存访问单元,所述存储单元包括:寄存器、缓存中任意组合;

所述缓存,用于存储所述第一输入数据以及所述第二输入数据;

所述寄存器,用于存储所述第一输入数据以及所述第二输入数据中标量数据;

所述缓存包括高速暂存缓存;

所述控制器单元包括:指令存储单元、指令存储单元和存储队列单元;

所述指令存储单元,用于存储人工神经网络运算关联的计算指令;

所述指令处理单元,用于对所述计算指令解析得到多个运算指令;

所述存储队列单元,用于存储指令队列,该指令队列包括:按该队列的前后顺序待执行的多个运算指令或计算指令;

所述控制单元包括主处理电路,所述主处理电路包括:依赖关系处理单元;

所述依赖关系处理单元,用于确定第一运算指令与所述第一运算指令之前的第零运算指令是否存在关联关系,如所述第一运算指令与所述第零运算指令存在关联关系,将所述第一运算指令缓存在所述指令存储单元内,在所述第零运算指令执行完毕后,从所述指令存储单元提取所述第一运算指令传输至所述运算单元;

所述确定该第一运算指令与第一运算指令之前的第零运算指令是否存在关联关系包括:

依据所述第一运算指令提取所述第一运算指令中所需数据的第一存储地址区间,依据所述第零运算指令提取所述第零运算指令中所需数据的第零存储地址区间,如所述第一存储地址区间与所述第零存储地址区间具有重叠的区域,确定所述第一运算指令与所述第零运算指令具有关联关系,如所述第一存储地址区间与所述第零存储地址区间不具有重叠的

区域，确定所述第一运算指令与所述第零运算指令不具有关联关系。

10.一种机器学习运算装置，其特征在于，所述机器学习运算装置包括一个或多个如权利要求1-9任一项所述的计算装置，用于从其他处理装置中获取待运算输入数据和控制信息，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计算装置时，所述多个所述计算装置间可以通过特定的结构进行连接并传输数据；

其中，多个所述计算装置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；多个所述计算装置共享同一控制系统或拥有各自的控制系统；多个所述计算装置共享内存或者拥有各自的内存；多个所述计算装置的互联方式是任意互联拓扑。

11.一种组合处理装置，其特征在于，所述组合处理装置包括如权利要求10所述的机器学习运算装置，通用互联接口、存储装置和其他处理装置；

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；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。

12.一种神经网络芯片，其特征在于，所述机器学习芯片包括如权利要求10所述的机器学习运算装置或如权利要求11所述的组合处理装置或如权利要求11所述的组合处理装置。

13.一种电子设备，其特征在于，所述电子设备包括如所述权利要求12所述的芯片。

14.一种板卡，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12所述的神经网络芯片；

其中，所述神经网络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，用于对所述芯片的状态进行监控。

15.一种执行机器学习模型的计算方法，其特征在于，所述计算方法应用于计算装置，所述计算装置用于执行机器学习计算；所述计算装置包括：压缩单元、运算单元以及控制器单元；所述方法包括：

所述控制器单元获取针对第一输入数据的压缩请求，并根据所述压缩请求指示所述压缩单元对所述第一输入数据进行压缩；其中，所述第一输入数据包括第一权值矩阵；

所述压缩单元，用于将第一权值矩阵压缩为第二权值矩阵；

所述控制器单元获取第二输入数据以及计算指令；所述第二输入数据包括所述第二权值矩阵以及输入神经元数据；

所述控制器单元解析该计算指令得到多个运算指令，将所述多个运算指令以及所述第二输入数据发送给运算单元；

所述运算单元获取所述运算指令，并根据所述运算指令以及所述第二输入数据执行神经网络计算。

16.根据权利要求15所述的方法，其特征在于，所述压缩单元包括：分解单元、求解单元以及训练单元；

其中，所述分解单元，用于将所述第一权值矩阵分解成第三权值矩阵；其中，所述第三

权值矩阵包括至少两个子矩阵；

所述求解单元，用于根据第一公式确定所述至少两个子矩阵中的每个子矩阵的大小，所述第一公式为$Q \approx Q_1 * Q_2 * \ldots\ldots * Q_n$；其中，所述Q表示第一权值矩阵；所述$Q_1$表示所述至少两个子矩阵中的第一子矩阵；所述$Q_2$表示所述至少两个子矩阵中的第二子矩阵；所述$Q_n$表示所述至少两个子矩阵中的第n子矩阵；

所述训练单元，用于调整所述至少两个子矩阵中的每个子矩阵的大小，并通过训练压缩后的机器学习模型，以得到满足预设精度的第二权值矩阵。

17.根据权利要求16所述的方法，其特征在于，所述求解单元，用于根据第一公式确定所述至少两个子矩阵中的每个子矩阵，所述第一公式为$Q \approx Q_1 * Q_2 * \ldots\ldots * Q_n$，包括：

所述求解单元，具体用于根据所述第一公式和第二公式确定所述至少两个子矩阵中的每个子矩阵的大小，所述第二公式为$||Q - Q_1 * Q_2 * \ldots\ldots * Q_n|| \leqslant T$，其中，所述T表示预设的误差阈值。

18.根据权利要求16所述的方法，其特征在于，所述训练单元，用于调整所述至少两个子矩阵中的每个子矩阵的大小，并通过训练压缩后的机器学习模型，以得到满足预设精度的第二权值矩阵，包括：

所述训练单元，具体用于调整所述至少两个子矩阵中的每个子矩阵的大小，并通过训练压缩后的机器学习模型，以得到满足预设精度并且与所述第一权值矩阵之间的压缩比满足预设压缩比的第二权值矩阵。

19.根据权利要求16-18任一项所述的方法，其特征在于，所述计算装置用于执行全连接层神经网络计算；所述至少两个子矩阵包括两个子矩阵；所述第一公式包括：$M \approx M_1 * M_2$；所述两个子矩阵包括第一子矩阵M1和第二子矩阵M2，所述M1为$N_{in} * K$矩阵，所述M2为$K * N_{out}$矩阵；其中，K为压缩参数，N_{in}为所述神经网络的输入神经元的个数，N_{out}为所述神经网络的输出神经元的个数；所述压缩参数用于表征所述M1的输出神经元的个数以及所述M2的输入神经元的个数，所述K为大于0且小于等于$\min(N_{in}, N_{out})$的正整数。

20.根据权利要求16-18任一项所述的方法，其特征在于，所述计算装置用于执行卷积层神经网络计算；所述卷积层神经网络包括$N_{fin} * N_{fout}$个卷积核；所述第一公式包括：$F \approx F_1 * F_2$；其中，F表示所述$N_{fin} * N_{fout}$个卷积核中的任意一个卷积核；所述F1为第一子卷积核；所述F2为第二子卷积核；所述第一子卷积核F1为(K_x, R)，所述第二子卷积核F2为(R, K_y)，(K_x, K_y)表示卷积核的大小，R为压缩参数，所述R为大于0且小于等于$\min(K_x, K_y)$的正整数。

21.根据权利要求16-18任一项所述的方法，其特征在于，所述计算装置用于执行LSTM层神经网络计算，所述LSTM层包括N个全连接层，所述N为大于0的正整数；针对第j个全连接层，所述第一公式包括：$M_j \approx M_{j\_1} * M_{j\_2}$；所述第j个全连接层中的两个子矩阵包括第一子矩阵$M_{j_1}$和第二子矩阵$M_{j_2}$，所述$M_{j_1}$为$N_{in_j} * S$矩阵，所述$M_{j_2}$为$S * N_{out_j}$矩阵；其中，S为压缩参数，$N_{in_j}$为所述神经网络第j个全连接层的输入神经元的个数，$N_{out_j}$为所述神经网络第j个全连接层的输出神经元的个数；所述压缩参数用于表征所述M_{j_1}的输出神经元的个数以及所述M_{j_2}的输入神经元的个数，所述S为大于0且小于等于$\min(N_{in_j}, N_{out_j})$的正整数。

22.根据权利要求15所述的方法，其特征在于，所述运算单元包括：一个主处理电路和多个从处理电路；

所述主处理电路对所述第二输入数据执行前序处理以及与所述多个从处理电路之间

传输数据和运算指令；

所述多个从处理电路根据从所述主处理电路传输的数据以及运算指令并行执行中间运算得到多个中间结果，并将所述多个中间结果传输给所述主处理电路；

所述主处理电路对所述多个中间结果执行后续处理得到所述计算指令的计算结果。

23.根据权利要求15所述的方法，其特征在于，所述计算装置还包括：存储单元和直接内存访问单元，所述存储单元包括：寄存器、缓存中任意组合；

所述缓存存储所述第一输入数据以及所述第二输入数据；

所述寄存器存储所述第一输入数据以及所述第二输入数据中的标量；所述缓存包括高速暂存缓存；

所述控制器单元包括：指令存储单元、指令存储单元和存储队列单元；

所述指令存储单元存储人工神经网络运算关联的计算指令；

所述指令处理单元对所述计算指令解析得到多个运算指令；

所述存储队列单元存储指令队列，该指令队列包括：按该队列的前后顺序待执行的多个运算指令或计算指令；

所述控制单元包括主处理电路，所述主处理电路包括：依赖关系处理单元；

所述依赖关系处理单元确定第一运算指令与所述第一运算指令之前的第零运算指令是否存在关联关系，如所述第一运算指令与所述第零运算指令存在关联关系，将所述第一运算指令缓存在所述指令存储单元内，在所述第零运算指令执行完毕后，从所述指令存储单元提取所述第一运算指令传输至所述运算单元；

所述确定该第一运算指令与第一运算指令之前的第零运算指令是否存在关联关系包括：

依据所述第一运算指令提取所述第一运算指令中所需数据的第一存储地址区间，依据所述第零运算指令提取所述第零运算指令中所需数据的第零存储地址区间，如所述第一存储地址区间与所述第零存储地址区间具有重叠的区域，确定所述第一运算指令与所述第零运算指令具有关联关系，如所述第一存储地址区间与所述第零存储地址区间不具有重叠的区域，确定所述第一运算指令与所述第零运算指令不具有关联关系。

一种计算装置及相关产品

技术领域

[0001] 本申请涉及信息处理技术领域，具体涉及一种计算装置及相关产品。

背景技术

[0002] 神经网络是一种模仿动物神经网络行为特征，进行分布式并行信息处理的算法数学模型，这种网络由大量的节点(或称神经元)之间星湖连接构成，通过调整内部大量节点之间相互连接的关系，利用输入神经元数据、权值产生输出数据模拟人脑的信息处理过程处理信息并生成模式识别之后的结果。

[0003] 目前，神经网络被广泛应用在计算机视觉的各个领域，如图像识别、物体检测、图像分割等。然而，在实际应用中，神经网络模型往往有着数量庞大的模型参数(例如，超大规模权值)，在这种情况下，这意味着神经网络需要大量的计算资源和存储资源，大量的计算资源和存储资源的开销会降低神经网络的运算速度，对硬件的传输带宽以及运算器的要求也大大提高了，因此，如何在减少神经网络模型的参数的同时，降低神经网络的计算量变得十分重要。

[0004] 现有技术中，通过剪枝方法对神经网络模型的参数进行调整，以减少神经网络模型的参数以及降低神经网络的计算量。以对神经网络的权值进行剪枝为例，如图1A所示，在对神经网络的权值进行剪枝之前，神经网络的拓扑结构是规则的，然而，在对神经网络的权值进行剪枝之后，容易导致神经网络模型中原有的规则的拓扑结构变得不规则。那么，如何避免神经网络模型中的拓扑结构变得不规则是亟需解决的技术问题。

发明内容

[0005] 本申请实施例提供了一种计算装置及相关产品，在神经网络压缩过程中，可以保证神经网络模型的拓扑结构保持不变，从而避免了神经网络模型的拓扑结构出现不规则，减少了神经网络的运算量。

[0006] 第一方面，提供一种计算装置，所述计算装置用于执行机器学习模型机器学习计算，所述计算装置包括：压缩单元、运算单元以及控制器单元；

[0007] 所述控制器单元，用于获取针对第一输入数据的压缩请求，并根据所述压缩请求指示所述压缩单元对所述第一输入数据进行压缩；其中，所述第一输入数据包括第一权值矩阵；

[0008] 所述压缩单元，用于将第一权值矩阵压缩为第二权值矩阵；

[0009] 所述控制器单元，还用于获取第二输入数据以及计算指令；所述第二输入数据包括所述第二权值矩阵以及输入神经元数据；

[0010] 所述控制器单元，还用于解析该计算指令得到多个运算指令，将所述多个运算指令以及所述第二输入数据发送给运算单元；

[0011] 所述运算单元获取所述运算指令，并根据所述运算指令以及所述第二输入数据执行神经网络计算。

[0012] 通过本申请，可以通过压缩单元将第一权值矩阵压缩得到第二权值矩阵，继而可以根据第二权值矩阵以及输入神经元数据执行神经网络计算，解决了现有技术中采用神经网络剪枝算法容易带来的神经网络的拓扑结构出现不规则的情形，可以对神经网络进行深度压缩，可以减少神经网络的计算量，提高运算速度。

[0013] 第二方面，本申请实施例提供了一种机器学习运算装置，该机器学习运算装置包括一个或者多个第一方面所述的计算装置。该机器学习运算装置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；

[0014] 当所述机器学习运算装置包含多个所述计算装置时，所述多个所述计算装置间可以通过特定的结构进行链接并传输数据；

[0015] 其中，多个所述计算装置通过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；多个所述计算装置共享同一控制系统或拥有各自的控制系统；多个所述计算装置共享内存或者拥有各自的内存；多个所述计算装置的互联方式是任意互联拓扑。

[0016] 第三方面，本申请实施例提供了一种组合处理装置，该组合处理装置包括如第三方面所述的机器学习处理装置、通用互联接口，和其他处理装置。该机器学习运算装置与上述其他处理装置进行交互，共同完成用户指定的操作。该组合处理装置还可以包括存储装置，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。

[0017] 第四方面，本申请实施例提供了一种神经网络芯片，该神经网络芯片包括上述第一方面所述的计算装置、上述第二方面所述的机器学习运算装置或者上述第三方面所述的组合处理装置。

[0018] 第五方面，本申请实施例提供了一种神经网络芯片封装结构，该神经网络芯片封装结构包括上述第四方面所述的神经网络芯片。

[0019] 第六方面，本申请实施例提供了一种板卡，该板卡包括上述第五方面所述的神经网络芯片封装结构。

[0020] 第七方面，本申请实施例提供了一种电子装置，该电子装置包括上述第六方面所述的神经网络芯片或者上述第六方面所述的板卡。

[0021] 第八方面，本申请实施例还提供一种执行机器学习模型的计算方法，所述计算方法应用于计算装置，计算装置用于执行机器学习计算；所述计算装置包括：压缩单元、运算单元以及控制器单元；所述方法包括：

[0022] 所述控制器单元获取针对第一输入数据的压缩请求，并根据所述压缩请求指示所述压缩单元对所述第一输入数据进行压缩；其中，所述第一输入数据包括第一权值矩阵；

[0023] 所述压缩单元，用于将第一权值矩阵压缩为第二权值矩阵；

[0024] 所述控制器单元，还用于获取第二输入数据以及计算指令；所述第二输入数据包括所述第二权值矩阵以及输入神经元数据；

[0025] 所述控制器单元，还用于解析该计算指令得到多个运算指令，将所述多个运算指令以及所述第二输入数据发送给运算单元；

[0026] 所述运算单元，用于获取所述运算指令，并根据所述运算指令以及所述第二输入数据执行神经网络计算。

[0027] 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。

[0028] 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。

附图说明

[0029] 为了更清楚地说明本申请实施例中的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0030] 图1A是本申请实施例提供的一种对神经网络剪枝的操作示意图；

[0031] 图1B是本申请实施例提供的一种计算装置的结构示意图；

[0032] 图2是本申请实施例提供的一种控制单元的结构示意图；

[0033] 图3是本申请实施例提供的一种神经网络运算方法的流程示意图；

[0034] 图4是本申请实施例提供的一种神经网络压缩方法的流程示意图；

[0035] 图5A是本申请实施例提供的一种神经网络架构的示意图；

[0036] 图5B是本申请实施例提供的一种全连接层权值矩阵的示意图；

[0037] 图5C是本申请实施例提供的一种对全连接层权值矩阵进行压缩的操作示意图；

[0038] 图5D是本申请实施例提供的一种卷积层中卷积核的结构示意图；

[0039] 图5E是本申请另一实施例提供的一种全连接层权值矩阵的示意图；

[0040] 图5F是本申请实施例提供的一种对LSTM层进行压缩的操作示意图；

[0041] 图6是本申请实施例提供的另一种计算装置的结构示意图；

[0042] 图7是本申请实施例提供的主处理电路的结构示意图；

[0043] 图8是本申请实施例提供的另一种计算装置的结构示意图；

[0044] 图9是本申请实施例提供的树型模块的结构示意图；

[0045] 图10是本申请实施例提供的又一种计算装置的结构图；

[0046] 图11是本申请实施例提供的还一种计算装置的结构图；

[0047] 图12是本申请实施例提供的另一种计算装置的结构图；

[0048] 图13是本申请实施例提供的一种组合处理装置的结构图；

[0049] 图14是本申请实施例提供的另一种组合处理装置的结构图；

[0050] 图15是本申请实施例提供的一种板卡的结构示意图

[0051] 图16是本申请实施例提供的一种神经网络压缩方法的流程示意图；

[0052] 图17A是本申请实施例提供的一种神经网络压缩装置的结构示意图；

[0053] 图17B是本申请实施例提供的一种压缩单元的结构示意图；

[0054] 图18是本申请实施例提供的一种电子设备的结构示意图。

具体实施方式

[0055] 下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

[0056] 本申请的说明书和权利要求书及所述附图中的术语“第一”、“第二”、“第三”和“第四”等是用于区别不同对象，而不是用于描述特定顺序。此外，术语“包括”和“具有”以及它们任何变形，意图在于覆盖不排他的包含。例如包含了一系列步骤或单元的过程、方法、系统、产品或设备没有限定于已列出的步骤或单元，而是可选地还包括没有列出的步骤或单元，或可选地还包括对于这些过程、方法、产品或设备固有的其它步骤或单元。

[0057] 在本文中提及“实施例”意味着，结合实施例描述的特定特征、结构或特性可以包含在本申请的至少一个实施例中。在说明书中的各个位置出现该短语并不一定均是指相同的实施例，也不是与其它实施例互斥的独立的或备选的实施例。本领域技术人员显式地和隐式地理解的是，本文所描述的实施例可以与其它实施例相结合。

[0058] 本申请提供了一种压缩元，用于将第一权值矩阵压缩为第二权值矩阵，解决了现有技术中采用神经网络剪枝算法容易带来的神经网络的拓扑结构出现不规则的情形。在实际应用中，上述压缩单元可以用于神经网络计算中，具体地，用于执行神经网络计算的计算装置中，下面结合图1B所示的计算装置对本发明进行介绍。

[0059] 参阅图1B，图1B是本发明实施例提供的一种计算装置的结构示意图，该计算装置用于执行机器学习计算，该计算装置包括：控制器单元11、运算单元12以及压缩单元13，其中，控制器单元11分别与运算单元12以及压缩单元13连接；

[0060] 其中，控制器单元11，用于获取针对第一输入数据的压缩请求，并根据所述压缩请求指示所述压缩单元对所述第一输入数据进行压缩；其中，所述第一输入数据包括第一权值矩阵；在一种可选方案中，该压缩请求可以通过数据输入输出单元进行触发的，该数据输入输出单元具体可以为一个或多个数据I/O接口或I/O引脚；

[0061] 所述压缩单元13，用于将所述第一权值矩阵压缩为第二权值矩阵；其中，第二权值矩阵中包括至少两个子矩阵；

[0062] 具体实现中，所述压缩单元13包括分解单元131、求解单元132以及训练单元133。其中，分解单元131，用于将所述第一权值矩阵分解成第三权值矩阵；其中，所述第三权值矩阵包括至少两个子矩阵；求解单元132，用于根据第一公式确定所述至少两个子矩阵中的每个子矩阵的大小，所述第一公式为$Q \approx Q_1 * Q_2 * \ldots\ldots * Q_n$；其中，所述Q表示第一权值矩阵；所述$Q_1$表示所述至少两个子矩阵中的第一子矩阵；所述$Q_2$表示所述至少两个子矩阵中的第二子矩阵；所述$Q_n$表示所述至少两个子矩阵中的第n子矩阵；训练单元133，用于调整所述至少两个子矩阵中的每个子矩阵的大小，并通过训练压缩后的机器学习模型，以得到满足预设精度的第二权值矩阵。

[0063] 所述控制器单元11，还用于获取第二输入数据以及计算指令；所述第二输入数据包括第二权值矩阵以及输入神经元数据；在一种可选方案中，具体的，获取第二输入数据以及计算指令方式可以通过数据输入输出单元得到，该数据输入输出单元具体可以为一个或多个数据I/O接口或I/O引脚。

[0064] 所述控制器单元11，还用于解析该计算指令得到多个运算指令，将所述多个运算指令以及所述第二输入数据发送给运算单元；

[0065] 所述运算单元12，用于获取所述运算指令，并根据所述运算指令以及所述第二输入数据执行神经网络计算。

[0066] 在其中一个实现方式中，考虑到上述计算装置中设置有“压缩指令”，在这种情况下，控制器单元11，用于获取第一输入数据以及压缩指令；其中，所述第一输入数据包括第一权值矩阵；在一种可选方案中，具体的，获取第一输入数据以及压缩指令方式可以通过数据输入输出单元得到，该数据输入输出单元具体可以为一个或多个数据I/O接口或I/O引脚。

[0067] 所述控制器单元11，还用于解析该压缩指令得到多个操作指令，将所述多个操作指令以及所述第一权值矩阵发送给压缩单元；

[0068] 所述压缩单元13，用于根据所述多个操作指令将所述第一权值矩阵压缩为第二权值矩阵；

[0069] 所述控制器单元11，还用于获取第二输入数据以及计算指令；所述第二输入数据包括所述第二权值矩阵以及输入神经元数据；在一种可选方案中，具体的，获取第二输入数据以及计算指令方式可以通过数据输入输出单元得到，该数据输入输出单元具体可以为一个或多个数据I/O接口或I/O引脚。

[0070] 所述控制器单元11，还用于解析该计算指令得到多个运算指令，将所述多个运算指令以及所述第二输入数据发送给运算单元；

[0071] 所述运算单元12，用于获取所述运算指令，并根据所述运算指令以及所述第二输入数据执行神经网络计算。

[0072] 具体实现中，所述运算单元12包括主处理电路101以及多个从处理电路102，所述主处理电路101，用于对所述第二输入数据执行前序处理以及与所述多个从处理电路之间传输数据以及运算指令；

[0073] 多个从处理电路102，用于依据从所述主处理电路传输的数据以及运算指令并行执行中间运算得到多个中间结果，并将多个中间结果传输给所述主处理电路；

[0074] 主处理电路101，用于对所述多个中间结果执行后续处理得到所述计算指令的计算结果。

[0075] 可选的，上述第二输入数据具体可以包括：第二权值矩阵以及输入神经元数据。上述计算结果具体可以为：神经网络运算的结果即输出神经元数据。

[0076] 在其中一个实施方式中，上述计算装置还可以包括：该存储单元10和直接内存访问单元50，存储单元10可以包括：寄存器、缓存中的一个或任意组合，具体的，所述缓存，用于存储所述计算指令；所述寄存器，用于存储所述输入数据和标量；所述缓存为高速暂存缓存。直接内存访问单元50用于从存储单元10读取或存储数据。

[0077] 本申请实施例中，如图2所示，该控制器单元11包括：指令缓存单元110、指令处理单元111、依赖关系处理单元112以及存储队列单元113；

[0078] 指令缓存单元110，用于存储所述人工神经网络运算关联的计算指令，在第零计算指令在被执行的过程中，同时将未被提交执行的其他指令缓存在所述指令缓存单元110中，当所述第零计算指令执行完之后，如果第一计算指令是指令缓存单元110中未被提交指令

中最早的一条指令，则所述第一计算指令将被提交，一旦提交，该指令进行的操作对装置状态的改变将无法撤销；

[0079] 所述指令处理单元111，用于从所述指令缓存单元获取所述计算指令，并对所述计算指令解析得到多个操作指令；

[0080] 所述依赖关系处理单元112，用于在具有多个操作指令时，确定第一操作指令与所述第一操作指令之前的第零操作指令是否存在关联关系，如所述第一操作指令与所述第零操作指令存在关联关系，则将所述第一操作指令存储到指令队列单元113内，在所述第零操作指令执行完毕后，所述第一操作指令与所述第零操作指令的关联关系解除，则从所述指令队列单元113中提取所述第一操作指令传输至所述运算单元；

[0081] 所述确定该第一操作指令与第一操作指令之前的第零操作指令是否存在关联关系包括：

[0082] 依据所述第一操作指令提取所述第一操作指令中所需数据(例如矩阵)的第一存储地址区间，依据所述第零操作指令提取所述第零操作指令中所需矩阵的第零存储地址区间，如所述第一存储地址区间与所述第零存储地址区间具有重叠的区域，则确定所述第一操作指令与所述第零操作指令具有关联关系，如所述第一存储地址区间与所述第零存储地址区间不具有重叠的区域，则确定所述第一操作指令与所述第零操作指令不具有关联关系。

[0083] 存储队列单元113，用于存储指令队列，该指令队列包括：按该队列的前后顺序待执行的多个操作指令或计算指令。

[0084] 本申请实施例中，如图2所示，所述指令处理单元111包括取指模块、译码模块以及指令队列，其中，所述取指模块，用于从所述指令缓存单元110中获取神经网络的计算指令；所述译码模块用于对所述取指模块获取的计算指令进行译码，得到神经网络的操作指令；所述指令队列用于对译码后得到的操作指令，按照待执行的先后顺序进行顺序存储。

[0085] 举例说明，在一个可选的技术方案中，主运算处理电路也可以包括一个控制器单元，该控制器单元可以包括主指令处理单元，具体用于将指令译码成微指令。当然在另一种可选方案中，从运算处理电路也可以包括另一个控制器单元，该另一个控制器单元包括从指令处理单元，具体用于接收并处理微指令。上述微指令可以为指令的下一级指令，该微指令可以通过对指令的拆分或解码后获得，能被进一步解码为各部件、各单元或各处理电路的控制信号。

[0086] 在一种可选方案中，该计算指令的结构可以如下表所示。

[0087] 表1

[0088]

操作码	寄存器或立即数	寄存器/立即数	...

[0089] 上表中的省略号表示可以包括多个寄存器或立即数。

[0090] 在另一种可选方案中，该计算指令可以包括：一个或多个操作域以及一个操作码。该计算指令可以包括神经网络运算指令，也可以包括上述所涉及的压缩指令。以神经网络运算指令为例，如表1所示，其中，寄存器号0、寄存器号1、寄存器号2、寄存器号3、寄存器号4可以为操作域。其中，每个寄存器号0、寄存器号1、寄存器号2、寄存器号3、寄存器号4可以是一个或者多个寄存器的号码。

[0091] 表2

[0092]

操作码	寄存器号0	寄存器号1	寄存器号2	寄存器号3	寄存器号4
COMPUTE	输入数据起始地址	输入数据长度	权值起始地址	权值长度	激活函数插值表地址
IO	数据外部存储其地址	数据长度	数据内部存储器地址		
NOP					
JUMP	目标地址				
MOVE	输入地址	数据大小	输出地址		

[0093] 上述寄存器可以为片外存储器，当然在实际应用中，也可以为片内存储器，用于存储数据，该数据具体可以为n维数据，n为大于等于1的整数，例如，n＝1时，为1维数据，即向量，如n＝2时，为2维数据，即矩阵，如n＝3或3以上时，为多维张量。

[0094] 在本发明实施例中，所述计算装置执行所述神经网络运算的过程如图3所示，包括：

[0095] 步骤S1、控制器单元接收压缩指令，将压缩指令译码解析为多个操作指令，并将多个操作指令发送给压缩单元。

[0096] 控制器单元从存储单元读取压缩指令之后，将压缩指令解析为操作指令，并将所述操作指令发送至压缩单元。具体的，控制器单元11中指令处理单元111的取指模块从指令缓存单元110中获取压缩指令，并将该指令传送至译码模块，所述译码模块对所述压缩指令进行译码，得到操作指令，并将所述操作指令根据预设指令规则拆分为操作码和各个不同的操作域，其中，操作码和操作域的组成与作用可参照前文，在此不再赘述。所述译码模块将译码后得到的操作指令传送至指令队列中进行顺序存储，在所述指令队列中，根据所述运操作指令的操作码和操作与获取该指令对应的待处理数据的数据地址，并将所述数据地址传送至依赖关系处理单元112中，依赖关系处理单元分析该指令与正在执行的指令是否存在关联关系，若存在，则将该操作指令存储到存储队列单元113中直至所述关联关系解除，若不存在关联关系，则将该操作指令发送至压缩单元中执行对应的操作。

[0097] S2、压缩单元接收控制单元发送的操作指令，并根据从存储单元中读取的第一权值矩阵进行压缩处理，以得到满足预设精度的第二权值矩阵。

[0098] 下面结合图4所示的本发明实施例提供的神经网络压缩方法的流程示意图，具体阐述本发明实施例是如何实现针对第一权值矩阵的压缩，以得到第二权值矩阵的，可以包括但不限于如下步骤：

[0099] 步骤S21、将所述第一权值矩阵分解成第三权值矩阵；其中，所述第三权值矩阵包括至少两个子矩阵。

[0100] 具体实现中，第一权值矩阵中的权值数据可以为任意实数。这里，权值数据是指神经网络层与层之间的连接值，也即神经元之间的信息传递强度。

[0101] 在其中一个实施方式中，第三权值矩阵中包括两个子矩阵，这两个子矩阵中的每个子矩阵均包括压缩参数K。这里，压缩参数K为未知数，也即，在对第一权值矩阵进行分解

时，可以确定第一权值矩阵可以分解为两个子矩阵，但是不确定这两个子矩阵的每个子矩阵的大小规模。

[0102] 在其中另一个实施方式中，第三权值矩阵中的子矩阵的数量为n个，这里，n为大于2的正整数。这n个子矩阵中包括的压缩参数K的数量为(n-1)个。以将第一权值矩阵分为三个子矩阵为例，待求解的压缩参数K可以包括K1以及K2。

[0103] 步骤S22、根据第一公式确定所述至少两个子矩阵中的每个子矩阵的大小，所述第一公式为$Q \approx Q_1 * Q_2 * \ldots\ldots * Q_n$；其中，所述Q表示第一权值矩阵；所述$Q_1$表示所述至少两个子矩阵中的第一子矩阵；所述$Q_2$表示所述至少两个子矩阵中的第二子矩阵；所述$Q_n$表示所述至少两个子矩阵中的第n子矩阵。

[0104] 具体实现中，第一公式中的运算符号“*”表示矩阵的乘法运算。

[0105] 在其中一个实施方式中，当第三权值矩阵中包括两个子矩阵时，第一公式可以表示为：

[0106] $$Q \approx Q_1 * Q_2 \quad (1.1)$$

[0107] 在其中另一个实施方式中，当第三权值矩阵中包括至少两个子矩阵时，第一公式可以表示为：

[0108] $$Q \approx Q_1 * Q_2 * \ldots\ldots * Q_n \quad (1.2)$$

[0109] 上述公式(1.2)中，n为大于2的正整数。

[0110] 具体实现中，根据所述第一公式和第二公式确定所述至少两个子矩阵中的每个子矩阵的大小，所述第二公式为$||Q-Q_1*Q_2*\ldots\ldots*Q_n|| \leqslant T$，其中，所述T表示预设的误差阈值。

[0111] 具体实现中，这里所涉及的预设的误差阈值可以为5%-10%之间。可以理解的是，设置的预设的误差阈值越小，根据第一公式以及第二公式确定的至少两个子矩阵可以更好的表示第一权值矩阵的属性特征。

[0112] 步骤S23、调整所述至少两个子矩阵中的每个子矩阵的大小，并通过训练压缩后的机器学习模型，以得到满足预设精度的第二权值矩阵。

[0113] 具体实现中，调整至少两个子矩阵的每个子矩阵的大小的过程，其实质是压缩参数K值的动态变化过程，以寻找最佳的压缩参数K。随着压缩参数K值发生变化，针对第一权值矩阵与第二权值矩阵之间的压缩比也会发生变化。

[0114] 以语音识别的应用场景为例，在某一段单词序列中，可能存在一些单词被错误地插入、删除或替换的情况。例如，对于包含N个单词的一段初始识别文字而言，如果有I个单词背插入、D个单词被删除以及E个文字被替换，那么，词错误率WER(Word Error Rate，WER)为：

[0115] $$WER = (I+D+E)/N \quad (1.3)$$

[0116] 其中，错误率WER通常用百分比表示。

[0117] 在采用神经网络模型识别该段单词序列时，可以得到该段单词序列的词错误率的检测精度。在本发明实施例中，这里所涉及的预设精度为压缩前的神经网络模型针对词错误率WER的检测精度。例如，该预设精度为70%。在一般情况下，压缩后的神经网络的错误率WER会变大，这意味着压缩后的神经网络的精度会变差。

[0118] 在发明实施例中，通过测量不同压缩比(压缩参数K值不同)所对应的神经网络模

型的词错误率的检测精度，以得到满足预设精度的第二权值矩阵。

[0119] 在一种优选的实施方式中，所述训练单元，用于调整所述至少两个子矩阵中的每个子矩阵的大小，并通过训练压缩后的机器学习模型，以得到满足预设精度的第二权值矩阵，包括：

[0120] 所述训练单元，具体用于调整所述至少两个子矩阵中的每个子矩阵的大小，并通过训练压缩后的机器学习模型，以得到满足预设精度并且与所述第一权值矩阵之间的压缩比满足预设压缩比的第二权值矩阵。

[0121] 可以理解的是，在该实施方式中，当前状态下的压缩参数K值不仅可以使得神经网络模型获得最优的压缩效果，还可以使得该压缩后的神经网络模型在检测词错误率WER时满足预设精度。在神经网络模型处于最优的压缩效果时，可以进一步地减少神经网络的运算量。

[0122] 以神经网络的全连接层为例，全连接层是指对n-1层和n层而言，n-1层的任意一个节点，都和n层的所有节点有连接。具体地，参见图5A，是本发明实施例提供的一种神经网络的一维全连接层的结构示意图，如图5A所示，该神经网络包括输入层、隐含层以及输出层，其中，输入层到隐含层之间的这一全连接层的二维参数矩阵为(3,4)，该二维参数矩阵(3,4)表示在输入层到隐含层之间的全连接层结构中，输入神经元的个数为3，输出神经元的个数为4，权值数量为12。具体实现中，这12个权值可以表示为4行3列的权值矩阵，其权值矩阵的表现形式可以如图5B所示。

[0123] 在全连接层神经网络中，所述第一公式包括：$M \approx M_1 * M_2$；所述两个子矩阵指包括第一子矩阵M1和第二子矩阵M2，所述M1为$N_{in} * K$矩阵，所述M2为$K * N_{out}$矩阵；其中，K为压缩参数，N_{in}为所述神经网络的输入神经元的个数，N_{out}为所述神经网络的输出神经元的个数；所述压缩参数用于表征所述M1的输出神经元的个数以及所述M2的输入神经元的个数，所述K为大于0且小于等于$\min(N_{in}, N_{out})$的正整数。

[0124] 如前所述，调整两个子矩阵的每个子矩阵的大小的过程，其实质是压缩参数K值的动态变化过程，以寻找最佳的压缩参数K。在实际应用中，可以采用二分查找的方式来确定全连接层神经网络中的压缩参数K值，从而得到满足预设精度的第二权值矩阵。在其中一个实施方式中，利用二分查找方式确定的压缩参数K可以使得第二权值矩阵满足预设精度。在其中另一个实施方式中，利用二分查找方式确定的压缩参数K可以使得第二权值矩阵满足预设精度的同时，第一权值矩阵与第二权值矩阵的压缩比满足预设压缩比，也即，针对该神经网络模型的压缩获得较优的压缩效果。

[0125] 具体实现中，压缩参数K值不同，也即基于多个不同压缩比对第一权值矩阵进行压缩，这里，在全连接层神经网络中，压缩比为$X = \frac{N_{in} * N_{out}}{N_{in} * K + K * N_{out}}$。

[0126] 接下来具体阐述如何采用二分查找的方式来确定压缩参数K值。首先，设定两个参数KL和KR。初始化情况下，令KL＝1，KR＝$\min(N_{in}, N_{out})$。在调整参数过程中K＝(KL+KR)/2。如果$M_1 * M_2$表示的第二权值矩阵导致压缩后的神经网络模型的精度下降X%(这里，X＝1～10等等)时，调整参数KL，使得KL＝K。如果$M_1 * M_2$表示的第二权值矩阵导致压缩后的神经网络模型满足预设精度，那么，调整KR，使得KR＝K，重复执行上述步骤，直至满足结束条件K＝KL或者K＝KR。

[0127] 以图5A中输入层到隐含层之间的这一全连接层为例，压缩参数K值为大于0且小于等于3的正整数。通过上述二分查找的方式确定压缩参数K＝2，也即，满足预设精度的第二权值矩阵中的第一子矩阵M1为(3,2)矩阵，第二子矩阵M2为(2,4)矩阵。具体地，针对图5A中输入层到隐含层之间的这一全连接层的压缩可以如图5C所示。

[0128] 在其中一个实施方式中，当第一公式的表现形式如公式(1.2)所示时，也即第三权值矩阵中的子矩阵的数量为n个，这里，n为大于2的正整数，此时，压缩参数K的数量为(n-1)个，可以表示为K1,K2,……,K_{n-1}。在实际应用中，可以采用自适应算法(例如，遗传算法)来确定全连接层神经网络中的(n-1)个压缩参数K值，从而得到满足预设精度和/或满足压缩效果的第二权值矩阵。接下来具体阐述是如何采用遗传算法来确定全连接层神经网络中的(n-1)个压缩参数K值的：

[0129] 步骤1：随机产生种群：设定种群的规模为P个，设置最大迭代次数T_{max}，例如，T_{max}＝100。在初始状态下，设置迭代次数计数器t＝0；交叉概率Pc＝A(例如，A＝0.4)，变异概率Pm＝B(例如，B＝0.6)，种群的矩阵每一行表示一个基因串个体，每一列表示个体的数目；这里，每一个个体是一组关于压缩参数K(例如，K_j)值的解；

[0130] 步骤2：计算种群中每个个体的适应度；这里，适应度是指该个体对应的第一权值矩阵与第二权值矩阵的压缩比和/或精度，其中，压缩比用于表征针对神经网络的压缩效果。

[0131] 步骤3：将选择算子作用于种群，把优化的个体直接遗传到下一代；

[0132] 步骤4：将交叉算子中作用于种群，对于任意两个个体，随机产生若干基因串的位置点，交换两个个体该位置上的值；

[0133] 步骤5：将变异算子作用于种群，对于任意个体，随机产生若干基因串的位置，然后改动这些位置上的值；这里，变异是指随机改变K_j的值；

[0134] 步骤6：保留每一代中适应度最高的个体，进入下一代；

[0135] 步骤7：判断是否达到最大迭代次数T_{max}，若t＝T_{max}，则输出具有最大适应度的个体，终止计算；否则，跳到步骤2继续执行。

[0136] 从而可以根据上述遗传算法来确定全连接层神经网络中的(n-1)个压缩参数K值。

[0137] 以神经网络的卷积层为例，如图5D所示，卷积层可以认为是一个四维矩阵(N_{fin},N_{fout},K_x,K_y)，其中，N_{fin}为输入特征图像的数量，N_{fout}为输出特征图像的数量，(K_x,K_y)为卷积层中卷积核的大小。

[0138] 在卷积层神经网络中，所述卷积层神经网络包括N_{fin}*N_{fout}个卷积核；所述第一公式包括：F≈F_1*F_2；其中，F表示所述N_{fin}*N_{fout}个卷积核中的任意一个卷积核；所述F1为第一子卷积核；所述F2为第二子卷积核；所述第一子卷积核F1为(K_x,R)，所述第二子卷积核F2为(R,K_y)，(K_x,K_y)表示卷积核的大小，R为压缩参数，所述R为大于0且小于等于min(K_x,K_y)的正整数。

[0139] 如前所述，调整两个子矩阵的每个子矩阵的大小的过程，其实质是压缩参数R值的动态变化过程，以寻找最佳的压缩参数R。在实际应用中，可以采用二分查找的方式来确定卷积层神经网络中的压缩参数R值，从而得到满足预设精度的第二权值矩阵。

[0140] 具体实现中，压缩参数R值不同，也即基于多个不同压缩比对第一权值矩阵进行压

缩，这里，在卷积层神经网络中，压缩比 $X=\frac{K_x*K_y}{K_x*R+R*K_y}$。

[0141] 在本发明实施例中，采用二分查找的方式确定压缩参数R值的实现过程参考前述文字描述，此处不多加赘述。

[0142] 例如，在图5D所示的卷积层神经网络结构中，该卷积层中包括4个卷积核，卷积核大小为3*3，其中，第1个卷积核中，$N_{fin}=4$，$N_{fout}=6$，压缩参数R值为大于0且小于等于4的正整数。通过上述二分查找的方式确定压缩参数R＝4，也即，满足预设精度的第1个卷积核中的第一子卷积核F1为(3,4)矩阵，第二子卷积核F2为(4,3)矩阵。在其中一个实施方式中，针对图5D所示的其它卷积核，可以采用与第1个卷积核相同的压缩方法，也可以采用与第1个卷积核不同的压缩方法，本发明实施例不作具体限定。

[0143] 在其中一个实施方式中，当第一公式的表现形式如公式(1.2)所示时，可以采用自适应算法(例如，遗传算法)确定卷积层神经网络中的压缩参数R值，其具体实现过程请参考前述描述，此处不多加赘述。

[0144] 以神经网络的长短时记忆LSTM层(LSTM，Long Short-term Memory)为例，LSTM层的权值由多个全连接层权值组成。假设LSTM层的权值由t个全连接层权值组成，t为大于0的正整数。例如，第j个全连接层权值分别为(N_{in_j}，N_{out_j})，其中，N_{in_j}表示第j个全连接层输入神经元个数，N_{out_j}表示第j个全连接层输出神经元个数，第j个全连接层的权值数量为$N_{in_j}*N_{out_j}$。

[0145] 在LSTM层神经网络中，所述LSTM层包括N个全连接层，所述N为大于0的正整数；针对第j个全连接层，所述第一公式包括：$M_j \approx M_{j\_1}*M_{j\_2}$；所述第j个全连接层中的两个子矩阵包括第一子矩阵$M_{j_1}$和第二子矩阵$M_{j_2}$，所述$M_{j_1}$为$N_{in_j}$*S矩阵，所述$M_{j_2}$为S*$N_{out_j}$矩阵；其中，S为压缩参数，$N_{in_j}$为所述神经网络第j个全连接层的输入神经元的个数，$N_{out_j}$为所述神经网络第j个全连接层的输出神经元的个数；所述压缩参数用于表征所述M_{j_1}的输出神经元的个数以及所述M_{j_2}的输入神经元的个数，所述S为大于0且小于等于min(N_{in_j}，N_{out_j})的正整数。

[0146] 如前所述，调整两个子矩阵的每个子矩阵的大小的过程，其实质是压缩参数S值的动态变化过程，以寻找最佳的压缩参数S。在实际应用中，可以采用二分查找的方式来确定卷积层神经网络中的压缩参数S值，从而得到满足预设精度的第二权值矩阵。

[0147] 具体实现中，针对第j个全连接层，压缩参数S值不同，也即基于多个不同压缩比对第一权值矩阵进行压缩，这里，在第j个全连接层中，压缩比 $X=\frac{N_{in\_j}*N_{out\_j}}{N_{in}*S+S*N_{out}}$。

[0148] 在本发明实施例中，采用二分查找的方式确定第j个全连接层中的压缩参数S值的实现过程参考前述文字描述，此处不多加赘述。

[0149] 以图5A所示的神经网络架构为例，该神经网络包括输入层、隐含层以及输出层，其中，输入层到隐含层之间为第1个全连接层，隐含层到输出层之间为第2个全连接层。针对输入层到隐含层之间的这一全连接层结构(也即，第1个全连接层)的具体阐述请参考前述描述，此处不多加赘述。由图5A可知，隐含层到输出层之间的这一全连接层的二维参数矩阵为(4,2)，该二维参数矩阵(4,2)表示在隐含层到输出层之间的全连接层结构中，输入神经元的个数为4，输出神经元的个数为2，权值数量为8。具体实现中，这8个权值可以表示为2行4

列权值矩阵，其权值矩阵的表现形式可以如图5E所示。那么，在这种情况下，压缩参数S值为大于0且小于等于2的正整数。通过二分查找的方式确定压缩参数S＝2，也即，在第2个全连接层中，满足预设精度的第二权值矩阵中的第一子矩阵M_{2_1}为(4,2)矩阵，第二子矩阵M_{2_2}为(2,2)矩阵。具体地，针对图5A中输入层但隐含层之间的这一全连接层以及隐含层到输出层之间的这一全连接层的压缩可以如图5F所示。

[0150] 在其中一个实施方式中，当第一公式的表现形式如公式(1.2)所示时，可以采用自适应算法(例如，遗传算法)确定LSTM层神经网络中的压缩参数S值，其具体实现过程请参考前述描述，此处不多加赘述。

[0151] 通过本发明实施例，控制器单元在获取到压缩指令后，将其进行解析可以得到多个操作指令，之后，将这多个操作指令以及第一权值矩阵发送给压缩单元，继而压缩元通过对第一权值矩阵进行分解，可以得到第二权值矩阵。具体实现中，第二权值矩阵包括至少两个子矩阵，继而通过调整这至少两个子矩阵中的每个子矩阵的大小，以及结合训练压缩后的机器学习模型，以得到满足预设精度的第二权值矩阵，解决了现有技术中采用神经网络剪枝算法容易带来的神经网络的拓扑结构出现不规则的情形。此外，对神经网络进行压缩，可以减少神经网络的运算量，进而提高运算速度。

[0152] S3、控制器单元获取第二输入数据以及计算指令，其中，第二输入数据包括第二权值矩阵及输入神经元数据。

[0153] S4、控制器单元将计算指令解析为运算指令，将运算指令以及第二输入数据发送给运算单元。

[0154] 具体实现中，针对控制器单元获取计算指令，并将计算指令进行解析，以得到多个运算指令的实现方式，请参考前述控制器单元获取压缩指令的文字描述，此处不多加赘述。

[0155] S5、运算单元接收控制器单元发送的运算指令，并根据运算指令以及第二输入数据执行神经网络计算。

[0156] 在实际应用中，这里所涉及的神经网络计算可以包括人工神经网络运算，也可以包括卷积神经网络运算等等。

[0157] 以人工神经网络运算为例，对于人工神经网络运算来说，如果该人工神经网络运算具有多层运算，多层运算的输入神经元和输出神经元并非是指整个神经网络的输入层中神经元和输出层中神经元，而是对于网络中任意相邻的两层，处于网络正向运算下层中的神经元即为输入神经元，处于网络正向运算上层中的神经元即为输出神经元。以卷积神经网络为例，设一个卷积神经网络有L层，K＝1,2,...,L-1，对于第K层和第K+1层来说，我们将第K层称为输入层，其中的神经元为所述输入神经元，第K+1层称为输出层，其中的神经元为所述输出神经元。即除最顶层外，每一层都可以作为输入层，其下一层为对应的输出层。

[0158] 具体实现中，对于神经网络中的运算可以为神经网络中的一层的运算，对于多层神经网络，其实现过程是，在正向运算中，当上一层人工神经网络执行完成之后，下一层的运算指令会将运算单元中计算出的输出神经元作为下一层的输入神经元进行运算(或者是对该输出神经元进行某些操作再作为下一层的输入神经元)，同时，将权值也替换为下一层的权值；在反向运算中，当上一层人工神经网络的反向运算执行完成后，下一层运算指令会将运算单元中计算出的输入神经元梯度作为下一层的输出神经元梯度进行运算(或者是对该输入神经元梯度进行某些操作再作为下一层的输出神经元梯度)，同时将权值替换为下

一层的权值。

[0159] 以完成神经网络的正向运算过程为例，首先，运算单元从存储单元中读取第二输入数据，其中，第二输入数据包括第二权值矩阵以及输入神经元数据。

[0160] 其次，主处理电路读取相对应的神经元数据，并将所述神经元数据按照指定顺序依次广播给各个从处理电路。在实际应用中，神经元数据可以只广播一次，从处理电路接收该数据后暂存到缓存或寄存器中，便于对其进行复用。此外，神经元数据也可以进行多次广播，从处理电路接收到数据之后直接使用，无需复用。在一种可能的实施方式中，主处理电路读取所述神经元数据之后，直接将神经元数据进行广播。

[0161] 之后，每个从处理电路将读入的神经元数据和第二权值矩阵根据运算指令进行内积运算，而后将内积结果传递回主处理电路。

[0162] 在其中一个实施方式中，从处理电路可以将每次执行内积运算得到的部分和传输回主处理电路进行累加；在在其中一个实施方式中，也可以将每次从处理电路执行的内积运算得到的部分和保存在从处理电路的寄存器和/或片上缓存中，累加结束之后传输回主处理电路；在其中一个实施方式中，也可以将每次从处理电路执行的内积运算得到的部分和在部分情况下保存在从处理电路的寄存器和/或片上缓存中进行累加，部分情况下传输到主处理电路进行累加，累加结束之后传输回主处理电路。

[0163] 最后，主处理电路将各从处理电路的结果进行累加、激活等操作后，直到完成神经网络的正向运算过程，得到预测结果和实际结果间的误差值，即最后一层的神经元梯度数据，保存到存储单元。

[0164] 在本发明实施例中，运算单元12可以设置成一主多从结构。在一种可选实施例中，运算单元12如图6所示，可以包括一个主处理电路101和多个从处理电路102。在一个实施例里，如图6所示，多个从处理电路呈阵列分布；每个从处理电路与相邻的其他从处理电路连接，主处理电路连接所述多个从处理电路中的k个从处理电路，所述k个从处理电路为：第1行的n个从处理电路、第m行的n个从处理电路以及第1列的m个从处理电路，需要说明的是，如图6所示的K个从处理电路仅包括第1行的n个从处理电路、第m行的n个从处理电路以及第1列的m个从处理电路，即该k个从处理电路为多个从处理电路中直接与主处理电路连接的从处理电路。

[0165] K个从处理电路，用于在所述主处理电路以及多个从处理电路之间的数据以及指令的转发。

[0166] 可选的，如图7所示，该主处理电路还可以包括：转换处理电路110、激活处理电路111、加法处理电路112中的一种或任意组合；

[0167] 转换处理电路110，用于将主处理电路接收的数据块或中间结果执行第一数据结构与第二数据结构之间的互换(例如连续数据与离散数据的转换)；或将主处理电路接收的数据块或中间结果执行第一数据类型与第二数据类型之间的互换(例如定点类型与浮点类型的转换)；

[0168] 激活处理电路111，用于执行主处理电路内数据的激活运算；

[0169] 加法处理电路112，用于执行加法运算或累加运算。

[0170] 所述主处理电路，用于将确定所述输入神经元为广播数据，权值为分发数据，将分发数据分配成多个数据块，将所述多个数据块中的至少一个数据块以及多个运算指令中的

至少一个运算指令发送给所述从处理电路；

[0171] 所述多个从处理电路，用于依据该运算指令对接收到的数据块执行运算得到中间结果，并将运算结果传输给所述主处理电路；

[0172] 所述主处理电路，用于将多个从处理电路发送的中间结果进行处理得到该计算指令的结果，将该计算指令的结果发送给所述控制器单元。

[0173] 所述从处理电路包括：乘法处理电路；

[0174] 所述乘法处理电路，用于对接收到的数据块执行乘积运算得到乘积结果；

[0175] 转发处理电路(可选的)，用于将接收到的数据块或乘积结果转发。

[0176] 累加处理电路，所述累加处理电路，用于对该乘积结果执行累加运算得到该中间结果。

[0177] 另一个实施例里，该运算指令为矩阵乘以矩阵的指令、累加指令、激活指令等等计算指令。

[0178] 下面通过神经网络运算指令来说明如图1B所示的计算装置的具体计算方法。对于神经网络运算指令来说，其实际需要执行的公式可以为：$s=s(\Sigma wx_i+b)$，其中，即将权值w乘以输入数据x_i，进行求和，然后加上偏置b后做激活运算s(h)，得到最终的输出结果s。

[0179] 在一种可选的实施方案中，如图8所示，所述运算单元包括：树型模块40，所述树型模块包括：一个根端口401和多个支端口404，所述树型模块的根端口连接所述主处理电路，所述树型模块的多个支端口分别连接多个从处理电路中的一个从处理电路；上述树型模块具有收发功能，所述树型模块具有收发功能，用于转发所述主处理电路与所述多个从处理电路之间的数据块、权值以及运算指令，即可以将主处理电路的数据传送给各个从处理电路，也可以将各个从处理电路的数据传送给主处理电路。

[0180] 可选的，该树型模块为计算装置的可选择结果，其可以包括至少1层节点，该节点为具有转发功能的线结构，该节点本身可以不具有计算功能。如树型模块具有零层节点，即无需该树型模块。

[0181] 可选的，该树型模块可以为n叉树结构，例如，如图9所示的二叉树结构，当然也可以为三叉树结构，该n可以为大于等于2的整数。本申请具体实施方式并不限制上述n的具体取值，上述层数也可以为2，从处理电路可以连接除倒数第二层节点以外的其他层的节点，例如可以连接如图9所示的倒数第一层的节点。

[0182] 可选的，上述运算单元可以携带单独的缓存，如图10所示，可以包括：神经元缓存单元，该神经元缓存单元63缓存该从处理电路的输入神经元向量数据和输出神经元值数据。

[0183] 如图11所示，该运算单元还可以包括：权值缓存单元64，用于缓存该从处理电路在计算过程中需要的权值数据。

[0184] 在一种可选实施例中，运算单元12如图12所示，可以包括分支处理电路103；其具体的连接结构如图12所示，其中，

[0185] 主处理电路101与分支处理电路103(一个或多个)连接，分支处理电路103与一个或多个从处理电路102连接；

[0186] 分支处理电路103，用于执行转发主处理电路101与从处理电路102之间的数据或指令。

[0187] 在一种可选实施例中，以神经网络运算中的全连接运算为例，过程可以为：y＝f(wx+b)，其中，x为输入神经元矩阵，w为权值矩阵，b为偏置标量，f为激活函数，具体可以为：sigmoid函数，tanh、relu、softmax函数中的任意一个。这里假设为二叉树结构，具有8个从处理电路，其实现的方法可以为：

[0188] 控制器单元从存储单元内获取输入神经元矩阵x，权值矩阵w以及全连接运算指令，将输入神经元矩阵x，权值矩阵w以及全连接运算指令传输给主处理电路；

[0189] 主处理电路确定该输入神经元矩阵x为广播数据，确定权值矩阵w为分发数据，将权值矩阵w拆分成8个子矩阵，然后将8个子矩阵通过树型模块分发给8个从处理电路，将输入神经元矩阵x广播给8个从处理电路，

[0190] 从处理电路并行执行8个子矩阵与输入神经元矩阵x的乘法运算和累加运算得到8个中间结果，将8个中间结果发送给主处理电路；

[0191] 主处理电路，用于将8个中间结果排序得到wx的运算结果，将该运算结果执行偏置b的运算后执行激活操作得到最终结果y，将最终结果y发送至控制器单元，控制器单元将该最终结果y输出或存储至存储单元内。

[0192] 如图1B所示的计算装置执行神经网络正向运算指令的方法具体可以为：

[0193] 控制器单元从指令存储单元内提取神经网络正向运算指令、神经网络运算指令对应的操作域以及至少一个操作码，控制器单元将该操作域传输至数据访问单元，将该至少一个操作码发送至运算单元。

[0194] 控制器单元从存储单元内提取该操作域对应的权值w和偏置b(当b为0时，不需要提取偏置b)，将权值w和偏置b传输至运算单元的主处理电路，控制器单元从存储单元内提取输入数据Xi，将该输入数据Xi发送至主处理电路。

[0195] 主处理电路依据该至少一个操作码确定为乘法运算，确定输入数据Xi为广播数据，确定权值数据为分发数据，将权值w拆分成n个数据块；

[0196] 控制器单元的指令处理单元依据该至少一个操作码确定乘法指令、偏置指令和累加指令，将乘法指令、偏置指令和累加指令发送至主处理电路，主处理电路将该乘法指令、输入数据Xi以广播的方式发送给多个从处理电路，将该n个数据块分发给该多个从处理电路(例如具有n个从处理电路，那么每个从处理电路发送一个数据块)；多个从处理电路，用于依据该乘法指令将该输入数据Xi与接收到的数据块执行乘法运算得到中间结果，将该中间结果发送至主处理电路，该主处理电路依据该累加指令将多个从处理电路发送的中间结果执行累加运算得到累加结果，依据该偏置指令将该累加结果执行加偏置b得到最终结果，将该最终结果发送至该控制器单元。

[0197] 另外，加法运算和乘法运算的顺序可以调换。

[0198] 本申请提供的技术方案通过一个指令即神经网络运算指令即实现了神经网络的乘法运算以及偏置运算，在神经网络计算的中间结果均无需存储或提取，减少了中间数据的存储以及提取操作，所以其具有减少对应的操作步骤，提高神经网络的计算效果的优点。

[0199] 本申请还揭露了一个机器学习运算装置，其包括一个或多个在本申请中提到的计算装置，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，执行结果通过I/O接口传递给外围设备。外围设备譬如摄像头，显示器，鼠标，键盘，网卡，wifi接口，服务器。当包含一个以上计算装置时，计算装置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，以支持更大规模的机器学习的运算。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。

[0200] 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。

[0201] 本申请还揭露了一个组合处理装置，其包括上述的机器学习运算装置，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。图13为组合处理装置的示意图。

[0202] 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。

[0203] 通用互联接口，用于在所述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。

[0204] 可选的，该结构如图14所示，还可以包括存储装置，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。存储装置用于保存在所述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。

[0205] 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。

[0206] 在一些实施例里，还申请了一种芯片，其包括了上述机器学习运算装置或组合处理装置。

[0207] 在一些实施例里，申请了一种芯片封装结构，其包括了上述芯片。

[0208] 在一些实施例里，申请了一种板卡，其包括了上述芯片封装结构。参阅图15，图15提供了一种板卡，上述板卡除了包括上述芯片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；

[0209] 所述存储器件390与所述芯片封装结构内的芯片通过总线连接，用于存储数据。所述存储器件可以包括多组存储单元393。每一组所述存储单元与所述芯片通过总线连接。可以理解，每一组所述存储单元可以是DDR SDRAM(英文：Double Data Rate SDRAM，双倍速率同步动态随机存储器)。

[0210] 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。在一个实施例中，所述存储装置可以包括4组所述存储单元。每一组所述存储单元可以包括多个DDR4颗粒(芯片)。在一个实施例中，所述芯片内部可以包括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,当每一组所述存储单元中采用DDR4-3200颗粒时,数据传输的理论带宽可达到25600MB/s。

[0211] 在一个实施例中,每一组所述存储单元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。在所述芯片中设置控制DDR的控制器,用于对每个所述存储单元的数据传输与数据存储的控制。

[0212] 所述接口装置与所述芯片封装结构内的芯片电连接。所述接口装置用于实现所述芯片与外部设备(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,所述接口装置可以为标准PCIE接口。比如,待处理的数据由服务器通过标准PCIE接口传递至所述芯片,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,所述接口装置还可以是其他的接口,本申请并不限制上述其他的接口的具体表现形式,所述接口单元能够实现转接功能即可。另外,所述芯片的计算结果仍由所述接口装置传送回外部设备(例如服务器)。

[0213] 所述控制器件与所述芯片电连接。所述控制器件用于对所述芯片的状态进行监控。具体的,所述芯片与所述控制器件可以通过SPI接口电连接。所述控制器件可以包括单片机(Micro Controller Unit,MCU)。如所述芯片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,所述芯片可以处于多负载和轻负载等不同的工作状态。通过所述控制装置可以实现对所述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。

[0214] 在一些实施例里,申请了一种电子设备,其包括了上述板卡。

[0215] 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。

[0216] 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。

[0217] 在本发明实施例中,考虑到针对神经网络的压缩方法可以包括但不限于应用在上述计算装置中,还可以应用在其它场景下,例如,减少神经网络的精度损失。基于此,下面结合图16所示的本发明实施例提供的神经网络压缩方法的流程示意图,具体说明本发明是如何实现针对第一权值矩阵的压缩,以得到第二权值矩阵的,可以包括但不限于如下步骤:

[0218] 步骤S100、获取第一输入数据;其中,所述第一输入数据包括第一权值矩阵。

[0219] 具体实现中,第一权值矩阵中的权值数据可以为任意实数。这里,权值数据是指神经网络层与层之间的连接值,也即神经元之间的信息传递强度。

[0220] 步骤S102、将所述第一权值矩阵压缩为第二权值矩阵;其中,第二权值矩阵中包括至少两个子矩阵。

[0221] 在其中一个实施方式中,所述将所述第一权值矩阵调整为第二权值矩阵,包括:

[0222] 将所述第一权值矩阵分解成第三权值矩阵;其中,所述第三权值矩阵包括至少两个子矩阵;

[0223] 根据第一公式确定所述至少两个子矩阵中的每个子矩阵的大小,所述第一公式为$Q \approx Q_1 * Q_2 * \ldots\ldots * Q_n$;其中,其中,所述Q表示第一权值矩阵;所述$Q_1$表示所述至少两个子矩

阵中的第一子矩阵;所述Q_2表示所述至少两个子矩阵中的第二子矩阵;所述Q_n表示所述至少两个子矩阵中的第n子矩阵;

[0224] 调整所述至少两个子矩阵中的每个子矩阵的大小,并通过训练压缩后的机器学习模型,以得到满足预设精度的第二权值矩阵。

[0225] 具体实现中,第一公式中的运算符号"*"表示矩阵的乘法运算。

[0226] 在其中一个实施方式中,当第三权值矩阵中包括两个子矩阵时,第一公式可以表示为:

[0227] $Q \approx Q_1 * Q_2$ (1.1)

[0228] 其中,Q表示第一权值矩阵,所述Q_1表示所述至少两个子矩阵中的第一子矩阵;所述Q_2表示所述至少两个子矩阵中的第二子矩阵。

[0229] 在其中另一个实施方式中,当第三权值矩阵中包括至少两个子矩阵时,第一公式可以表示为:

[0230] $Q \approx Q_1 * Q_2 * \ldots\ldots * Q_n$ (1.2)

[0231] 上述公式(1.2)中,n为大于2的正整数。

[0232] 具体实现中,将第一权值矩阵压缩为第二权值矩阵的实现过程中,当应用到不同的神经网络时(例如,全连接层神经网络、卷积层神经网络、LSTM层神经网络),上述所涉及的针对第一权值矩阵的分解操作、求解至少两个子矩阵中的每个子矩阵以及调整至少两个子矩阵中的每个子矩阵以获得满足预设精度的第二权值矩阵将有所差异,接下来将进行具体阐述:

[0233] (1)全连接层神经网络:

[0234] 全连接层是指对n-1层和n层而言,n-1层的任意一个节点,都和n层的所有节点有连接。具体地,参见图5A,是本发明实施例提供的一种神经网络的一维全连接层的结构示意图,如图5A所示,该神经网络包括输入层、隐含层以及输出层,其中,输入层到隐含层之间的这一全连接层的二维参数矩阵为(3,4),该二维参数矩阵(3,4)表示在输入层到隐含层之间的全连接层结构中,输入神经元的个数为3,输出神经元的个数为4,权值数量为12。具体实现中,这12个权值可以表示为4行3列的权值矩阵,其权值矩阵的表现形式可以如图5B所示。

[0235] 在全连接层神经网络中,所述第一公式包括:$M \approx M_1 * M_2$;所述两个子矩阵指包括第一子矩阵M1和第二子矩阵M2,所述M1为$N_{in} * K$矩阵,所述M2为$K * N_{out}$矩阵;其中,K为压缩参数,N_{in}为所述神经网络的输入神经元的个数,N_{out}为所述神经网络的输出神经元的个数;所述压缩参数用于表征所述M1的输出神经元的个数以及所述M2的输入神经元的个数,所述K为大于0且小于等于$\min(N_{in},N_{out})$的正整数。

[0236] 如前所述,调整两个子矩阵的每个子矩阵的大小的过程,其实质是压缩参数K值的动态变化过程,以寻找最佳的压缩参数K。在实际应用中,可以采用二分查找的方式来确定全连接层神经网络中的压缩参数K值,从而得到满足预设精度的第二权值矩阵。在其中一个实施方式中,利用二分查找方式确定的压缩参数K可以使得第二权值矩阵满足预设精度。在其中另一个实施方式中,利用二分查找方式确定的压缩参数K可以使得第二权值矩阵满足预设精度的同时,第一权值矩阵与第二权值矩阵的压缩比满足预设压缩比,也即,针对该神经网络模型的压缩获得较优的压缩效果。

[0237] 具体实现中,压缩参数K值不同,也即基于多个不同压缩比对第一权值矩阵进行压

缩，这里，在全连接层神经网络中，压缩比为$X=\frac{N_{in}*N_{out}}{N_{in}*K+K*N_{out}}$。

[0238] 接下来具体阐述如何采用二分查找的方式来确定压缩参数K值。首先，设定两个参数KL和KR。初始化情况下，令KL＝1，KR＝min(Nin，Nout)。在调整参数过程中K＝(KL+KR)/2。如果M1*M2表示的第二权值矩阵导致压缩后的神经网络模型的精度下降X％(这里，X＝1～10等等)时，调整参数KL，使得KL＝K。如果M1*M2表示的第二权值矩阵导致压缩后的神经网络模型满足预设精度，那么，调整KR，使得KR＝K，重复执行上述步骤，直至满足结束条件K＝KL或者K＝KR。

[0239] 以图5A中输入层到隐含层之间的这一全连接层为例，压缩参数K值为大于0且小于等于3的正整数。通过上述二分查找的方式确定压缩参数K＝2，也即，满足预设精度的第二权值矩阵中的第一子矩阵M1为(3,2)矩阵，第二子矩阵M2为(2,4)矩阵。具体地，针对图5A中输入层到隐含层之间的这一全连接层的压缩可以如图5C所示。

[0240] 在其中一个实施方式中，当第一公式的表现形式如公式(1.2)所示时，也即第三权值矩阵中的子矩阵的数量为n个，这里，n为大于2的正整数，此时，压缩参数K的数量为(n-1)个。在实际应用中，可以采用自适应算法(例如，遗传算法)来确定全连接层神经网络中的(n-1)个压缩参数K值，从而得到满足预设精度和/或满足压缩效果的第二权值矩阵。接下来具体阐述是如何采用遗传算法来确定全连接层神经网络中的(n-1)个压缩参数K值的：

[0241] 步骤1：随机产生种群：设定种群的规模为P个，设置最大迭代次数Tmax，例如，Tmax＝100。在初始状态下，设置迭代次数计数器t＝0；交叉概率Pc＝A(例如，A＝0.4)，变异概率Pm＝B(例如，B＝0.6)，种群的矩阵每一行表示一个基因串个体，每一列表示个体的数目；这里，每一个个体是一组关于压缩参数K(例如，Kj)值的解；

[0242] 步骤2：计算种群中每个个体的适应度；这里，适应度是指该个体对应的第一权值矩阵与第二权值矩阵的压缩比和/或精度，其中，压缩比用于表征针对神经网络的压缩效果。

[0243] 步骤3：将选择算子作用于种群，把优化的个体直接遗传到下一代；

[0244] 步骤4：将交叉算子中作用于种群，对于任意两个个体，随机产生若干基因串的位置点，交换两个个体该位置上的值；

[0245] 步骤5：将变异算子作用于种群，对于任意个体，随机产生若干基因串的位置，然后改动这些位置上的值；这里，变异是指随机改变Kj的值；

[0246] 步骤6：保留每一代中适应度最高的个体，进入下一代；

[0247] 步骤7：判断是否达到最大迭代次数Tmax，若t＝Tmax，则输出具有最大适应度的个体，终止计算；否则，跳到步骤2继续执行。

[0248] 从而可以根据上述遗传算法来确定全连接层神经网络中的(n-1)个压缩参数K值。

[0249] (2)卷积层神经网络：

[0250] 以神经网络的卷积层为例，如图5D所示，卷积层可以认为是一个四维矩阵(Nfin，Nfout，Kx，Ky)，其中，Nfin为输入特征图像的数量，Nfout为输出特征图像的数量，(Kx，Ky)为卷积层中卷积核的大小。

[0251] 在卷积层神经网络中，所述卷积层神经网络包括Nfin*Nfout个卷积核；所述第一公式包括：F≈F1*F2；其中，F表示所述Nfin*Nfout个卷积核中的任意一个卷积核；所述F1为第一

子卷积核;所述F2为第二子卷积核;所述第一子卷积核F1为(K_x,R),所述第二子卷积核F2为(R,K_y),(K_x,K_y)表示卷积核的大小,R为压缩参数,所述R为大于0且小于等于min(K_x,K_y)的正整数。

[0252] 如前所述,调整两个子矩阵的每个子矩阵的大小的过程,其实质是压缩参数R值的动态变化过程,以寻找最佳的压缩参数R。在实际应用中,可以采用二分查找的方式来确定卷积层神经网络中的压缩参数R值,从而得到满足预设精度的第二权值矩阵。

[0253] 具体实现中,压缩参数R值不同,也即基于多个不同压缩比对第一权值矩阵进行压缩,这里,在卷积层神经网络中,压缩比$X=\frac{K_x*K_y}{K_x*R+R*K_y}$。

[0254] 在本发明实施例中,采用二分查找的方式确定压缩参数R值的实现过程参考前述文字描述,此处不多加赘述。

[0255] 例如,图5D所示的卷积层神经网络结构中,该卷积层中包括4个卷积核,卷积核大小为3*3,其中,第1个卷积核中,N_{fin}=4,N_{fout}=6,压缩参数R值为大于0且小于等于4的正整数。通过上述二分查找的方式确定压缩参数R=4,也即,满足预设精度的第1个卷积核中的第一子卷积核F1为(3,4)矩阵,第二子卷积核F2为(4,3)矩阵。在其中一个实施方式中,针对图5D所示的其它卷积核,可以采用与第1个卷积核相同的压缩方法,也可以采用与第1个卷积核不同的压缩方法,本发明实施例不作具体限定。

[0256] 在其中一个实施方式中,当第一公式的表现形式如公式(1.2)所示时,可以采用自适应算法(例如,遗传算法)确定卷积层神经网络中的压缩参数R值,其具体实现过程请参考前述描述,此处不多加赘述。

[0257] (3)LSTM层神经网络:

[0258] 以神经网络的长短时记忆LSTM层(LSTM,Long Short-term Memory)为例,LSTM层的权值由多个全连接层权值组成。假设LSTM层的权值由t个全连接层权值组成,t为大于0的正整数。例如,第j个全连接层权值分别为(N_{in_j},N_{out_j}),其中,N_{in_j}表示第j个全连接层输入神经元个数,N_{out_j}表示第j个全连接层输出神经元个数,第j个全连接层的权值数量为N_{in_j}*N_{out_j}。

[0259] 在LSTM层神经网络中,所述LSTM层包括N个全连接层,所述N为大于0的正整数;针对第j个全连接层,所述第一公式包括:$M_j \approx M_{j\_1}*M_{j\_2}$;所述第j个全连接层中的两个子矩阵包括第一子矩阵$M_{j_1}$和第二子矩阵$M_{j_2}$,所述$M_{j_1}$为$N_{in_j}$*S矩阵,所述$M_{j_2}$为S*$N_{out_j}$矩阵;其中,S为压缩参数,$N_{in_j}$为所述神经网络第j个全连接层的输入神经元的个数,$N_{out_j}$为所述神经网络第j个全连接层的输出神经元的个数;所述压缩参数用于表征所述M_{j_1}的输出神经元的个数以及所述M_{j_2}的输入神经元的个数,所述S为大于0且小于等于min(N_{in_j},N_{out_j})的正整数。

[0260] 如前所述,调整两个子矩阵的每个子矩阵的大小的过程,其实质是压缩参数S值的动态变化过程,以寻找最佳的压缩参数S。在实际应用中,可以采用二分查找的方式来确定卷积层神经网络中的压缩参数S值,从而得到满足预设精度的第二权值矩阵。

[0261] 具体实现中,针对第j个全连接层,压缩参数S值不同,也即基于多个不同压缩比对第一权值矩阵进行压缩,这里,在第j个全连接层中,压缩比$X=\frac{N_{in_j}*N_{out_j}}{N_{in}*S+S*N_{out}}$。

[0262] 在本发明实施例中，采用二分查找的方式确定第j个全连接层中的压缩参数S值的实现过程参考前述文字描述，此处不多加赘述。

[0263] 以图5A所示的神经网络架构为例，该神经网络包括输入层、隐含层以及输出层，其中，输入层到隐含层之间为第1个全连接层，隐含层到输出层之间为第2个全连接层。针对输入层到隐含层之间的这一全连接层结构(也即，第1个全连接层)的具体阐述请参考前述描述，此处不多加赘述。由图5A可知，隐含层到输出层之间的这一全连接层的二维参数矩阵为(4,2)，该二维参数矩阵(4,2)表示在隐含层到输出层之间的全连接层结构中，输入神经元的个数为4，输出神经元的个数为2，权值数量为8。具体实现中，这8个权值可以表示为2行4列权值矩阵，其权值矩阵的表现形式可以如图5E所示。那么，在这种情况下，压缩参数S值为大于0且小于等于2的正整数。通过二分查找的方式确定压缩参数S＝2，也即，在第2个全连接层中，满足预设精度的第二权值矩阵中的第一子矩阵M_{2_1}为(4,2)矩阵，第二子矩阵M_{2_2}为(2,2)矩阵。具体地，针对图5A中输入层到隐含层之间的这一全连接层以及隐含层到输出层之间的这一全连接层的压缩可以如图5F所示。

[0264] 在其中一个实施方式中，当第一公式的表现形式如公式(1.2)所示时，可以采用自适应算法(例如，遗传算法)确定LSTM层神经网络中的压缩参数S值，其具体实现过程请参考前述描述，此处不多加赘述。

[0265] 步骤S104、根据第二输入数据执行神经网络计算，其中，第二输入数据包括第二权值矩阵以及神经元数据。

[0266] 在实际应用中，这里所涉及的神经网络计算可以包括人工神经网络运算，也可以包括卷积神经网络运算等等。

[0267] 以人工神经网络运算为例，对于人工神经网络运算来说，如果该人工神经网络运算具有多层运算，多层运算的输入神经元和输出神经元并非是指整个神经网络的输入层中神经元和输出层中神经元，而是对于网络中任意相邻的两层，处于网络正向运算下层中的神经元即为输入神经元，处于网络正向运算上层中的神经元即为输出神经元。以卷积神经网络为例，设一个卷积神经网络有L层，K＝1,2,...,L-1，对于第K层和第K+1层来说，我们将第K层称为输入层，其中的神经元为所述输入神经元，第K+1层称为输出层，其中的神经元为所述输出神经元。即除最顶层外，每一层都可以作为输入层，其下一层为对应的输出层。

[0268] 具体实现中，对于神经网络中的运算可以为神经网络中的一层的运算，对于多层神经网络，其实现过程是，在正向运算中，当上一层人工神经网络执行完成之后，下一层的运算指令会将运算单元中计算出的输出神经元作为下一层的输入神经元进行运算(或者是对该输出神经元进行某些操作再作为下一层的输入神经元)，同时，将权值也替换为下一层的权值；在反向运算中，当上一层人工神经网络的反向运算执行完成后，下一层运算指令会将运算单元中计算出的输入神经元梯度作为下一层的输出神经元梯度进行运算(或者是对该输入神经元梯度进行某些操作再作为下一层的输出神经元梯度)，同时将权值替换为下一层的权值。

[0269] 本发明实施例通过对第一权值矩阵进行分解，可以得到包含压缩参数的至少两个子矩阵，之后，根据公式求解这至少两个子矩阵中的每个子矩阵，并通过训练压缩后的神经网络以获得满足预设精度的第二权值矩阵，解决了现有技术中采用神经网络剪枝算法容易带来的神经网络的拓扑结构出现不规则的情形，可以对神经网络进行深度压缩，可以减少

神经网络的计算量，提高运算速度。

[0270] 为了便于更好地实施本发明实施例的上述方案，本发明还对应提供了一种神经网络压缩装置，下面结合附图来进行详细说明：

[0271] 如图17A所示的本发明实施例提供的神经网络压缩装置的结构示意图，该神经网络压缩装置包括：获取单元300、压缩单元13以及计算单元304；

[0272] 其中，所述获取单元300，用于获取第一输入数据；其中，所述第一输入数据包括第一权值矩阵；

[0273] 所述压缩单元13，用于将所述第一权值矩阵压缩为第二权值矩阵；其中，所述第二权值矩阵中包括至少两个子矩阵；

[0274] 所述计算单元304，用于根据第二输入数据执行神经网络计算，其中，所述第二输入数据包括所述第二权值矩阵以及输入神经元数据。

[0275] 在其中一个实施方式中，如图17B所示，压缩单元13包括分解单元130、求解单元131以及训练单元132；

[0276] 其中，所述分解单元130，用于将所述第一权值矩阵分解成第三权值矩阵；其中，所述第三权值矩阵包括至少两个子矩阵；

[0277] 所述求解单元131，用于根据第一公式确定所述至少两个子矩阵中的每个子矩阵的大小，所述第一公式为$Q \approx Q_1 * Q_2 * \ldots\ldots * Q_n$；其中，所述Q表示第一权值矩阵；所述$Q_1$表示所述至少两个子矩阵中的第一子矩阵；所述$Q_2$表示所述至少两个子矩阵中的第二子矩阵；所述$Q_n$表示所述至少两个子矩阵中的第n子矩阵；

[0278] 所述训练单元132，用于调整所述至少两个子矩阵中的每个子矩阵的大小，并通过训练压缩后的机器学习模型，以得到满足预设精度的第二权值矩阵。

[0279] 可选的，所述求解单元131，具体用于根据所述第一公式和第二公式确定所述至少两个子矩阵中的每个子矩阵的大小，所述第二公式为$||Q - Q_1 * Q_2 * \ldots\ldots * Q_n|| \leqslant T$，其中，所述T表示预设的误差阈值。

[0280] 可选的，所述训练单元132，具体用于调整所述至少两个子矩阵中的每个子矩阵的大小，并通过训练压缩后的机器学习模型，以得到满足预设精度并且与所述第一权值矩阵之间的压缩比满足预设压缩比的第二权值矩阵。

[0281] 可选的，所述神经网络包括全连接层神经网络；所述第一公式包括：$M \approx M_1 * M_2$；所述两个子矩阵指包括第一子矩阵M1和第二子矩阵M2，所述M1为$N_{in} * K$矩阵，所述M2为$K * N_{out}$矩阵；其中，K为压缩参数，N_{in}为所述神经网络的输入神经元的个数，N_{out}为所述神经网络的输出神经元的个数；所述压缩参数用于表征所述M1的输出神经元的个数以及所述M2的输入神经元的个数，所述K为大于0且小于等于$\min(N_{in}, N_{out})$的正整数。

[0282] 可选的，所述神经网络包括卷积层神经网络；所述卷积层神经网络包括$N_{fin} * N_{fout}$个卷积核；所述第一公式包括：$F \approx F_1 * F_2$；其中，F表示所述$N_{fin} * N_{fout}$个卷积核中的任意一个卷积核；所述F1为第一子卷积核；所述F2为第二子卷积核；所述第一子卷积核F1为(K_x, R)，所述第二子卷积核F2为(R, K_y)，(K_x, K_y)表示卷积核的大小，R为压缩参数，所述R为大于0且小于等于$\min(K_x, K_y)$的正整数。

[0283] 可选的，所述神经网络包括LSTM层神经网络，所述LSTM层包括N个全连接层，所述N为大于0的正整数；针对第j个全连接层，所述第一公式包括：$M_j \approx M_{j\_1} * M_{j\_2}$；所述第j个全连

接层中的两个子矩阵包括第一子矩阵M_{j_1}和第二子矩阵M_{j_2}，所述M_{j_1}为N_{in_j}*S矩阵，所述M_{j_2}为S*N_{out_j}矩阵；其中，S为压缩参数，N_{in_j}为所述神经网络第j个全连接层的输入神经元的个数，N_{out_j}为所述神经网络第j个全连接层的输出神经元的个数；所述压缩参数用于表征所述M_{j_1}的输出神经元的个数以及所述M_{j_2}的输入神经元的个数，所述S为大于0且小于等于$\min(N_{in_j},N_{out_j})$的正整数。

[0284] 本发明实施例通过对第一权值矩阵进行分解，可以得到包含压缩参数的至少两个子矩阵，之后，根据公式求解这至少两个子矩阵中的每个子矩阵，并通过训练压缩后的神经网络以获得满足预设精度的第二权值矩阵，解决了现有技术中采用神经网络剪枝算法容易带来的神经网络的拓扑结构出现不规则的情形，对神经网络进行深度压缩，可以减少神经网络的计算量，提高运算速度。

[0285] 为了便于更好地实施本发明实施例的上述方案，本发明还对应提供了另一种电子设备，下面结合附图来进行详细说明：

[0286] 如图18示出的本发明实施例提供的电子设备的结构示意图，电子设备40可以包括处理器401、存储器404和通信模块405，处理器401、存储器404和通信模块405可以通过总线406相互连接。存储器404可以是高速随机存储记忆体(Random Access Memory，RAM)存储器，也可以是非易失性的存储器(non-volatile memory)，例如至少一个磁盘存储器。存储器404可选的还可以是至少一个位于远离前述处理器401的存储系统。存储器404用于存储应用程序代码，可以包括操作系统、网络通信模块、用户接口模块以及数据处理程序，通信模块405用于与外部设备进行信息交互；处理器401被配置用于调用该程序代码，执行以下步骤：

[0287] 获取第一输入数据；其中，所述第一输入数据包括第一权值矩阵；

[0288] 将所述第一权值矩阵压缩为第二权值矩阵；其中，所述第二权值矩阵中包括至少两个子矩阵；

[0289] 根据第二输入数据执行神经网络计算，其中，所述第二输入数据包括所述第二权值矩阵以及输入神经元数据。

[0290] 其中，处理器401将所述第一权值矩阵压缩为第二权值矩阵；其中，所述第二权值矩阵中包括至少两个子矩阵，可以包括：

[0291] 将所述第一权值矩阵分解成第三权值矩阵；其中，所述第三权值矩阵包括至少两个子矩阵；

[0292] 确定所述至少两个子矩阵中的每个子矩阵的大小，所述第一公式为$Q\approx Q_1*Q_2*.....*Q_n$；其中，所述Q表示第一权值矩阵；所述$Q_1$表示所述至少两个子矩阵中的第一子矩阵；所述$Q_2$表示所述至少两个子矩阵中的第二子矩阵；所述$Q_n$表示所述至少两个子矩阵中的第n子矩阵；

[0293] 调整所述至少两个子矩阵中的每个子矩阵的大小，并通过训练压缩后的机器学习模型，以得到满足预设精度的第二权值矩阵。

[0294] 其中，处理器401根据第一公式确定所述至少两个子矩阵中的每个子矩阵，所述第一公式为$Q\approx Q_1*Q_2*.....*Q_n$，可以包括：

[0295] 根据所述第一公式和第二公式确定所述两个子矩阵中的每个子矩阵的大小，所述第二公式为$||Q-Q_1*Q_2*.....*Q_n||\leqslant T$，其中，所述T表示预设的误差阈值。

[0296] 其中，处理器401调整所述至少两个子矩阵中的每个子矩阵的大小，并通过训练压缩后的机器学习模型，以得到满足预设精度的第二权值矩阵，可以包括：

[0297] 调整所述至少两个子矩阵中的每个子矩阵的大小，并通过训练压缩后的机器学习模型，以得到满足预设精度并且与所述第一权值矩阵之间的压缩比满足预设压缩比的第二权值矩阵。

[0298] 其中，所述神经网络为全连接层神经网络；所述第一公式包括：$M \approx M_1 * M_2$；所述两个子矩阵指包括第一子矩阵M1和第二子矩阵M2，所述M1为$N_{in} * K$矩阵，所述M2为$K * N_{out}$矩阵；其中，K为压缩参数，N_{in}为所述神经网络的输入神经元的个数，N_{out}为所述神经网络的输出神经元的个数；所述压缩参数用于表征所述M1的输出神经元的个数以及所述M2的输入神经元的个数，所述K为大于0且小于等于$\min(N_{in}, N_{out})$的正整数。

[0299] 其中，所述神经网络为卷积层神经网络；所述卷积层神经网络包括$N_{fin} * N_{fout}$个卷积核；所述第一公式包括：$F \approx F_1 * F_2$；其中，F表示所述$N_{fin} * N_{fout}$个卷积核中的任意一个卷积核；所述F1为第一子卷积核；所述F2为第二子卷积核；所述第一子卷积核F1为(K_x, R)，所述第二子卷积核F2为(R, K_y)，(K_x, K_y)表示卷积核的大小，R为压缩参数，所述R为大于0且小于等于$\min(K_x, K_y)$的正整数。

[0300] 其中，所述神经网络为LSTM层神经网络；所述LSTM层神经网络包括N个全连接层，所述N为大于0的正整数；针对第j个全连接层，所述第一公式包括：$M_j \approx M_{j\_1} * M_{j\_2}$；所述第j个全连接层中的两个子矩阵包括第一子矩阵$M_{j_1}$和第二子矩阵$M_{j_2}$，所述$M_{j_1}$为$N_{in_j} * S$矩阵，所述$M_{j_2}$为$S * N_{out_j}$矩阵；其中，S为压缩参数，$N_{in_j}$为所述神经网络第j个全连接层的输入神经元的个数，$N_{out_j}$为所述神经网络第j个全连接层的输出神经元的个数；所述压缩参数用于表征所述M_{j_1}的输出神经元的个数以及所述M_{j_2}的输入神经元的个数，所述S为大于0且小于等于$\min(N_{in_j}, N_{out_j})$的正整数。

[0301] 需要说明的是，本发明实施例中的电子设备40中处理器的执行步骤可参考上述各方法实施例中图16实施例中的电子设备运行的具体实现方式，这里不再赘述。

[0302] 在实际应用中，电子设备40中的处理器401包括但不限于只有一个。在其中一个实施方式中，电子设备40中还包括处理图像的图形处理器GPU(GPU，Graphic Processing Uni)，也还可以包括嵌入式神经网络处理器(NPU，Neural-network Process Units)。此时，针对神经网络的压缩方法可以被集成在NPU中。在其中一个实施方式中，处理器401可以控制NPU执行针对第一权值矩阵的压缩方法。

[0303] 在具体实现中，如前所述，电子设备40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，本发明实施例不作具体限定。

[0304] 本发明实施例还提供了一种计算机存储介质，用于存储为上述图16所示的电子设备所用的计算机软件指令，其包含用于执行上述方法实施例所涉及的程序。通过执行存储的程序，可以实现针对第一权值矩阵的压缩，以得到满足预设精度的第二权值矩阵，从而避免了神经网络模型的拓扑结构出现不规则，减少了神经网络的运算量。

[0305] 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申请所必须的。

[0306] 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。

[0307] 在本申请所提供的几个实施例中，应该理解到，所揭露的装置，可通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如所述单元的划分，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，例如多个单元或组件可以结合或者可以集成到另一个系统，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，装置或单元的间接耦合或通信连接，可以是电性或其它的形式。

[0308] 所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元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单元来实现本实施例方案的目的。

[0309] 另外，在本申请各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元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。

[0310] 所述集成的单元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，本申请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)执行本申请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。

[0311] 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。

[0312] 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。

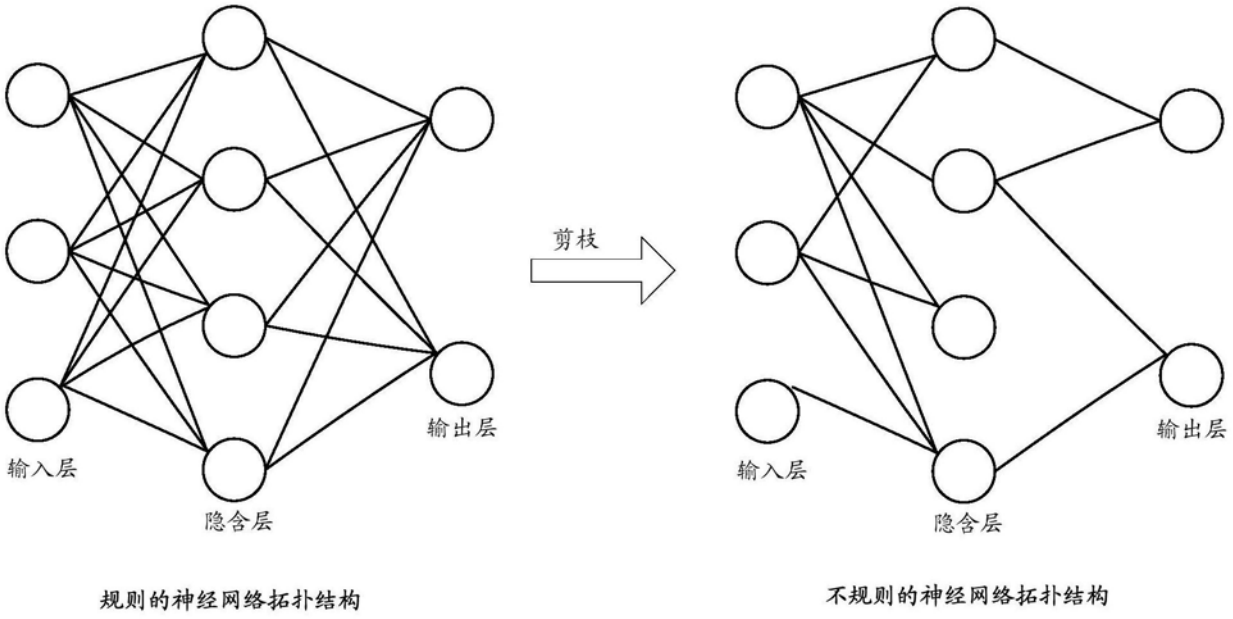


图1A

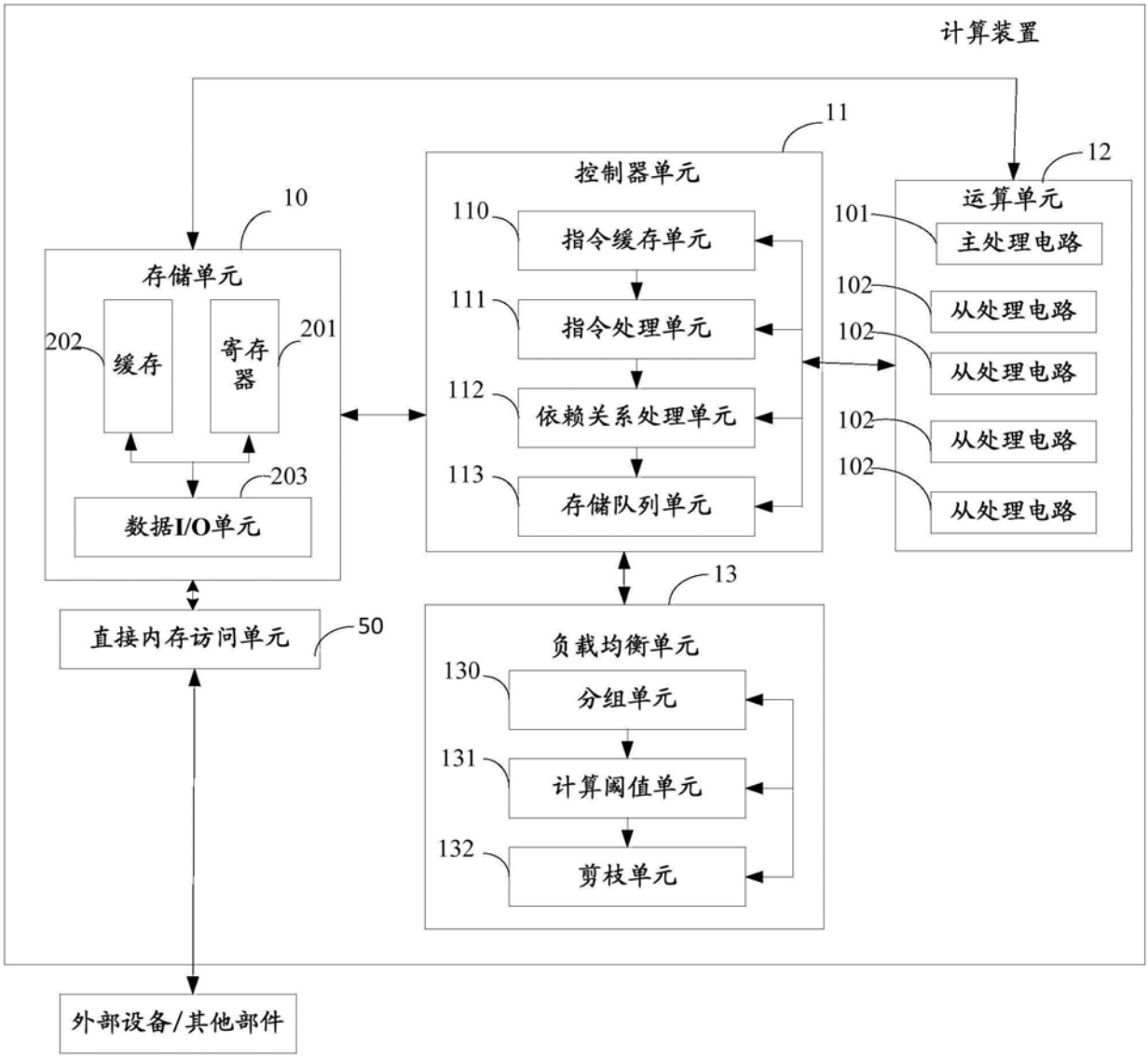

计算装置
11
控制器单元
12
运算单元
10
存储单元
110
指令缓存单元
101
主处理电路
111
指令处理单元
102
从处理电路
202
缓存
201
寄存器
112
依赖关系处理单元
102
从处理电路
102
从处理电路
113
存储队列单元
102
从处理电路
203
数据I/O单元
13
直接内存访问单元
50
负载均衡单元
130
分组单元
131
计算阈值单元
132
剪枝单元
外部设备/其他部件

图1B

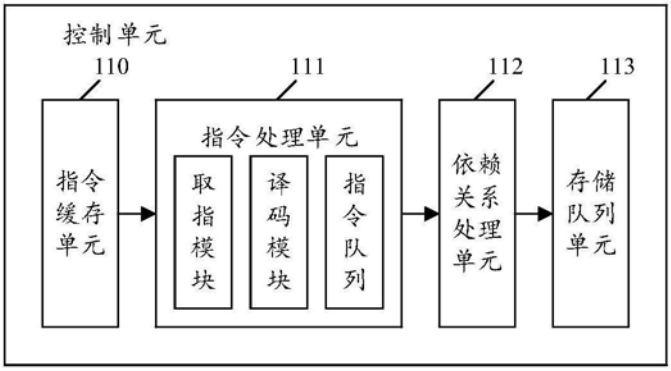

控制单元
110
111
112
113
指令缓存单元
指令处理单元
取指模块
译码模块
指令队列
依赖关系处理单元
存储队列单元

图2

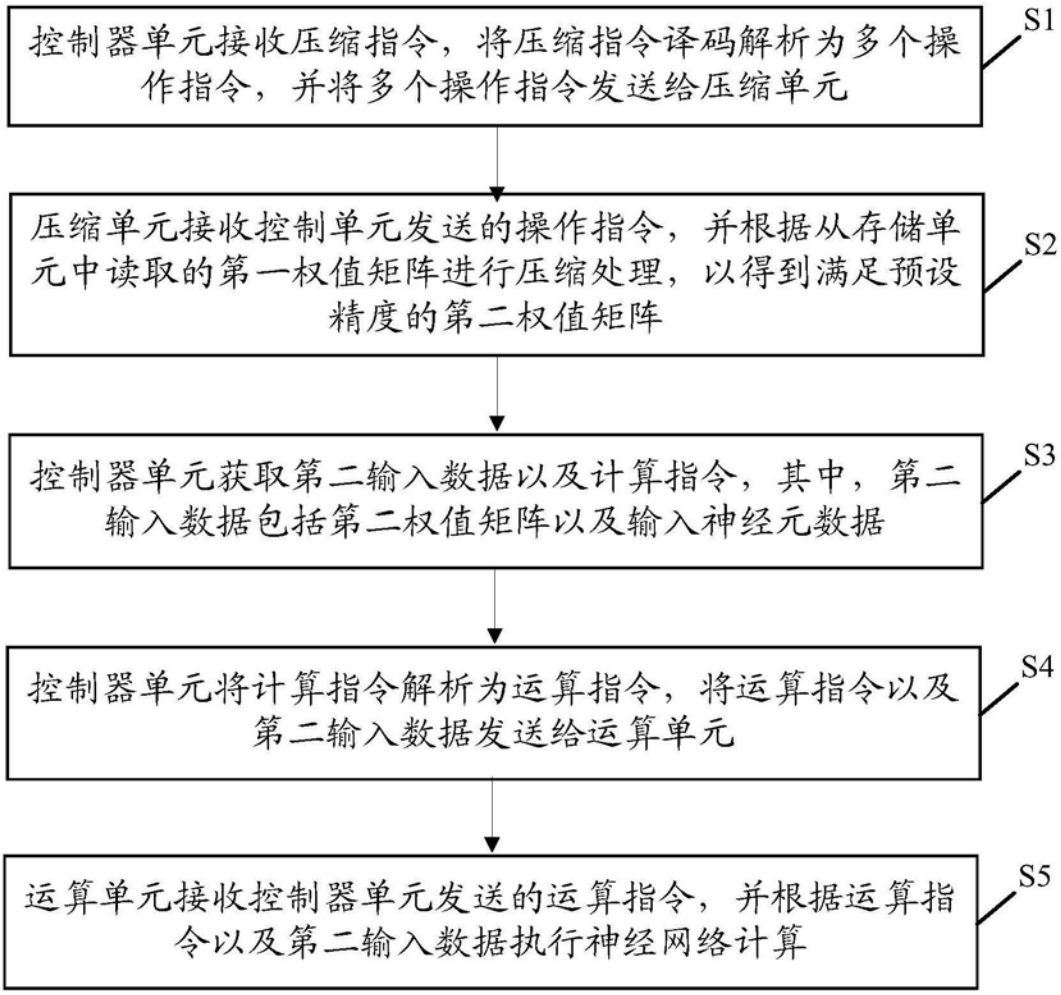

控制器单元接收压缩指令，将压缩指令译码解析为多个操作指令，并将多个操作指令发送给压缩单元
S1
压缩单元接收控制单元发送的操作指令，并根据从存储单元中读取的第一权值矩阵进行压缩处理，以得到满足预设精度的第二权值矩阵
S2
控制器单元获取第二输入数据以及计算指令，其中，第二输入数据包括第二权值矩阵以及输入神经元数据
S3
控制器单元将计算指令解析为运算指令，将运算指令以及第二输入数据发送给运算单元
S4
运算单元接收控制器单元发送的运算指令，并根据运算指令以及第二输入数据执行神经网络计算
S5

图3

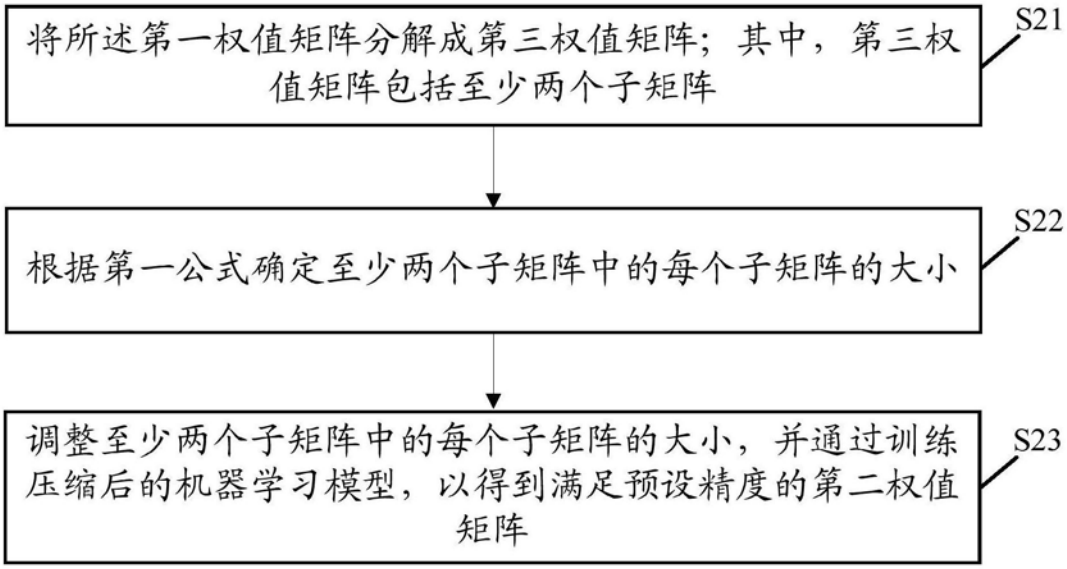

将所述第一权值矩阵分解成第三权值矩阵；其中，第三权值矩阵包括至少两个子矩阵
S21
根据第一公式确定至少两个子矩阵中的每个子矩阵的大小
S22
调整至少两个子矩阵中的每个子矩阵的大小，并通过训练压缩后的机器学习模型，以得到满足预设精度的第二权值矩阵
S23

图4

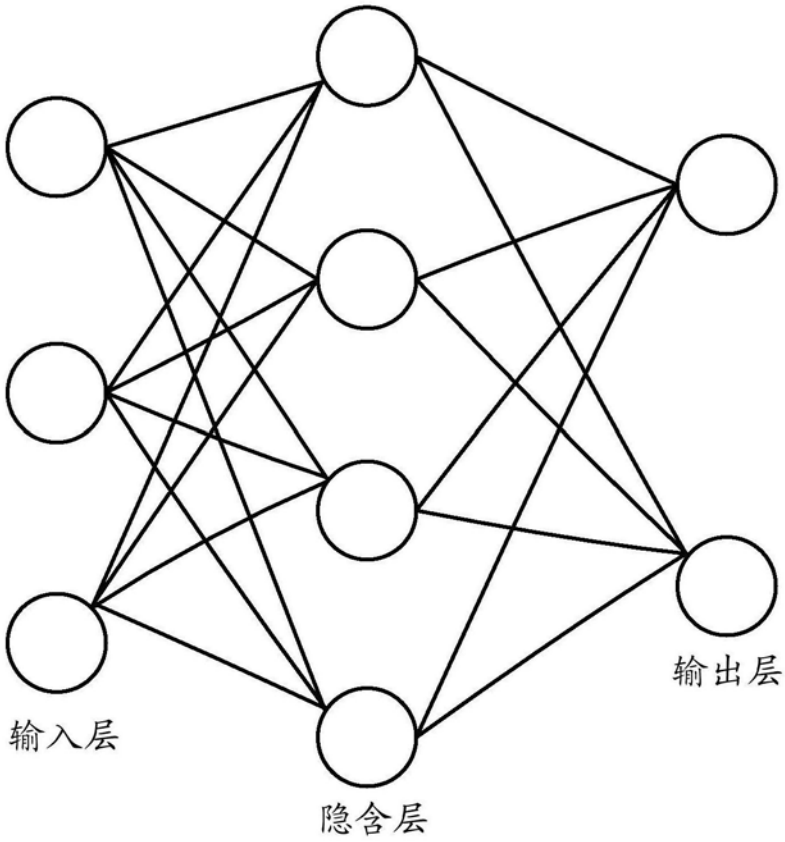


图5A

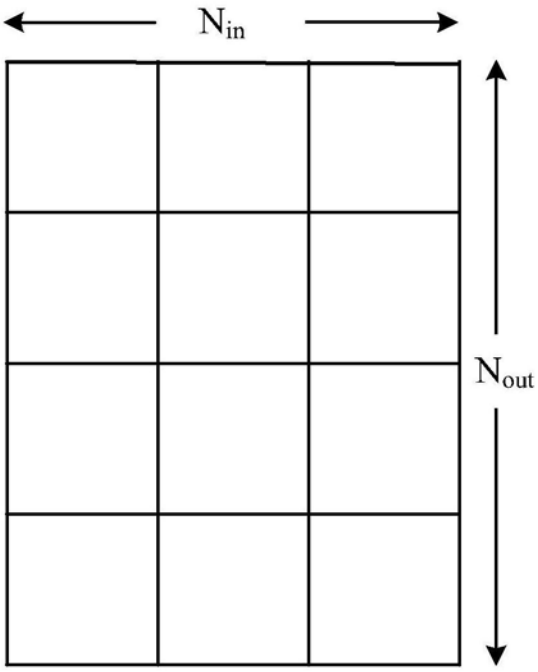


图5B

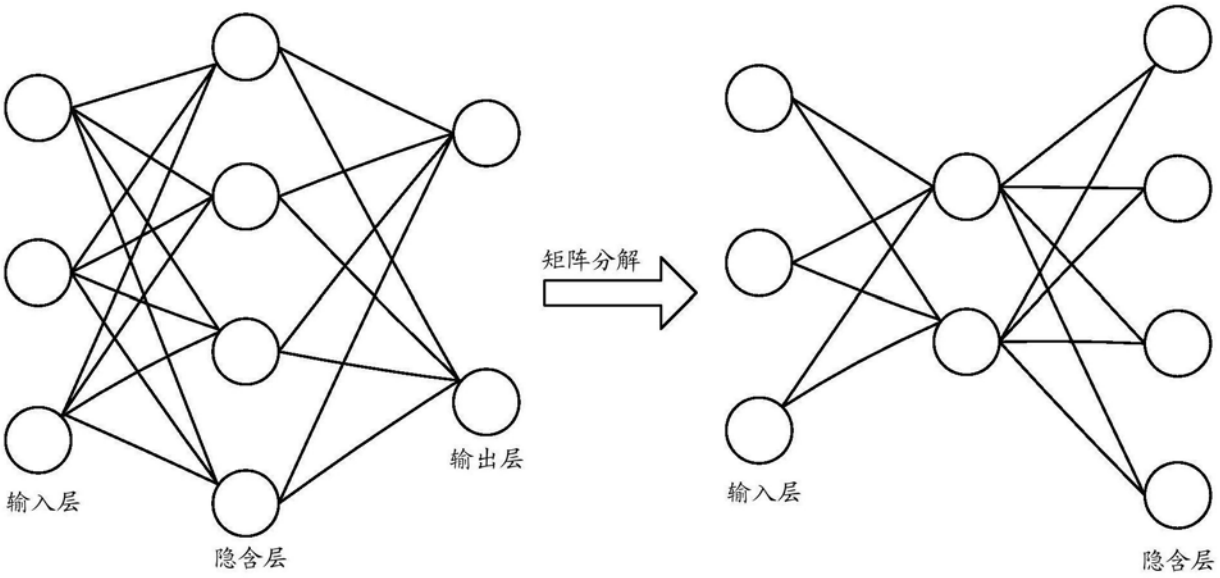


图5C

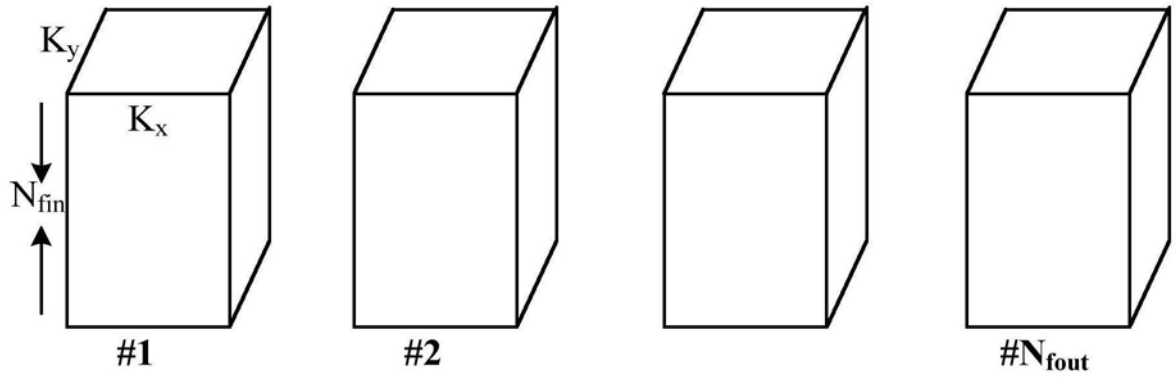


图5D

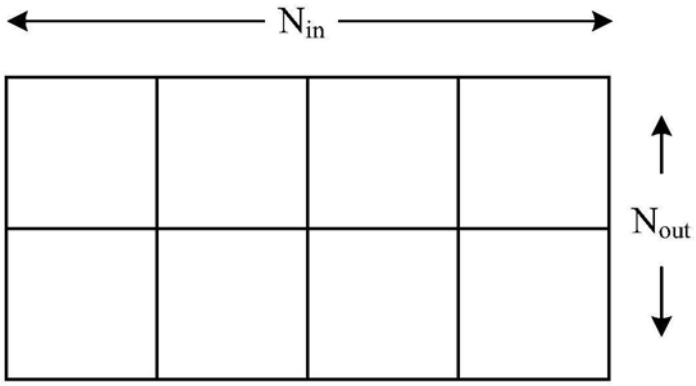


图5E

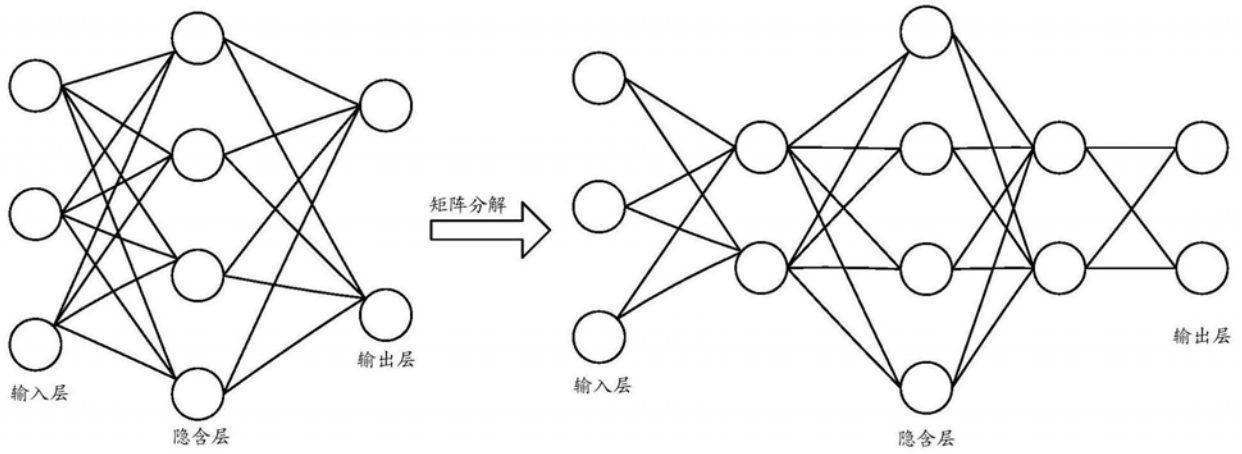


图5F

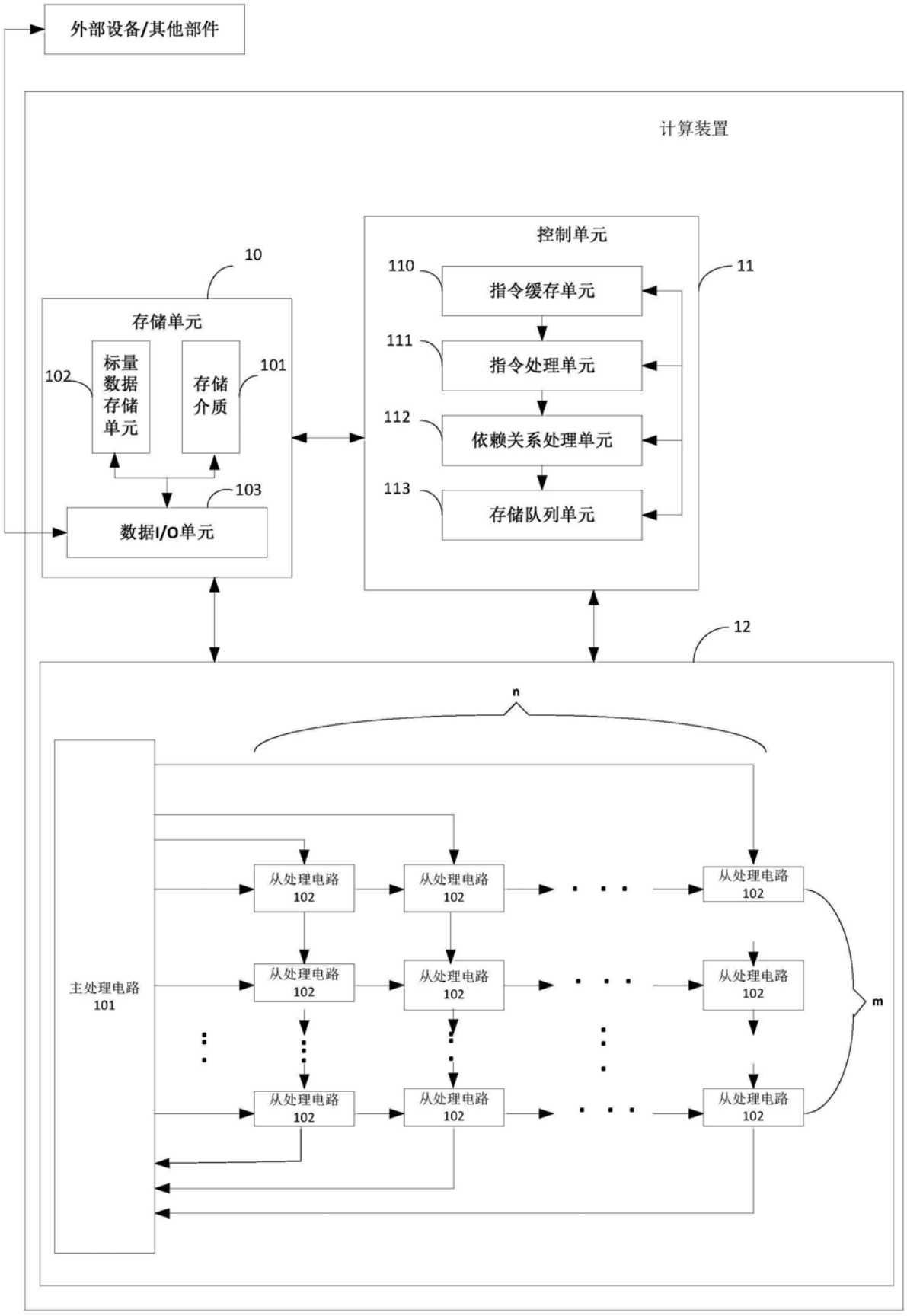

外部设备/其他部件
计算装置
控制单元
10
存储单元
标量数据存储单元
102
存储介质
101
103
数据I/O单元
110
指令缓存单元
11
111
指令处理单元
112
依赖关系处理单元
113
存储队列单元
12
n
主处理电路
101
从处理电路
102
m

图6


主处理电路101
转换处理电路110
激活处理电路111
加法处理电路112

图7

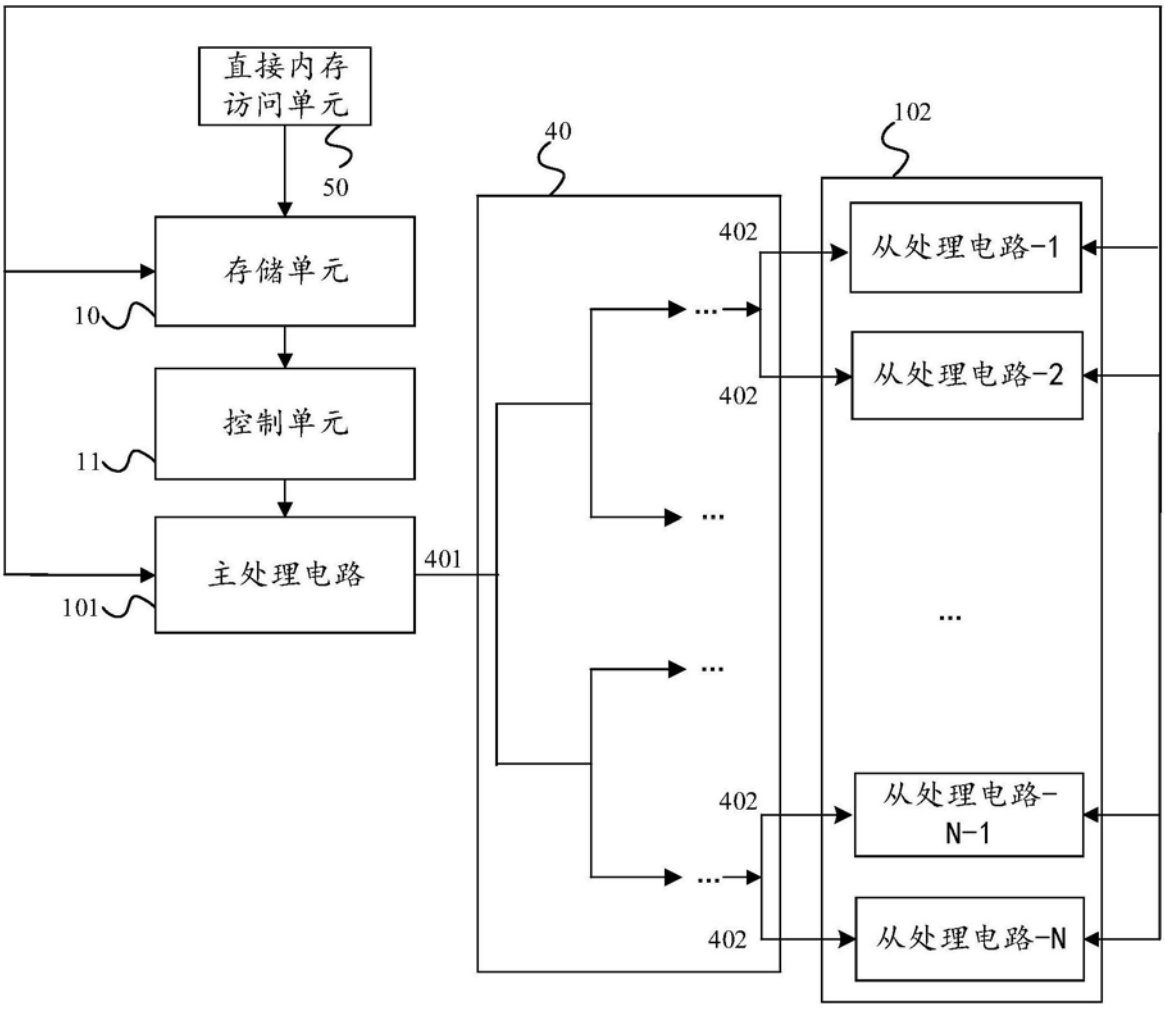

直接内存访问单元
50
存储单元
10
控制单元
11
主处理电路
101
401
40
402
402
402
402
102
从处理电路-1
从处理电路-2
从处理电路-N-1
从处理电路-N

图8

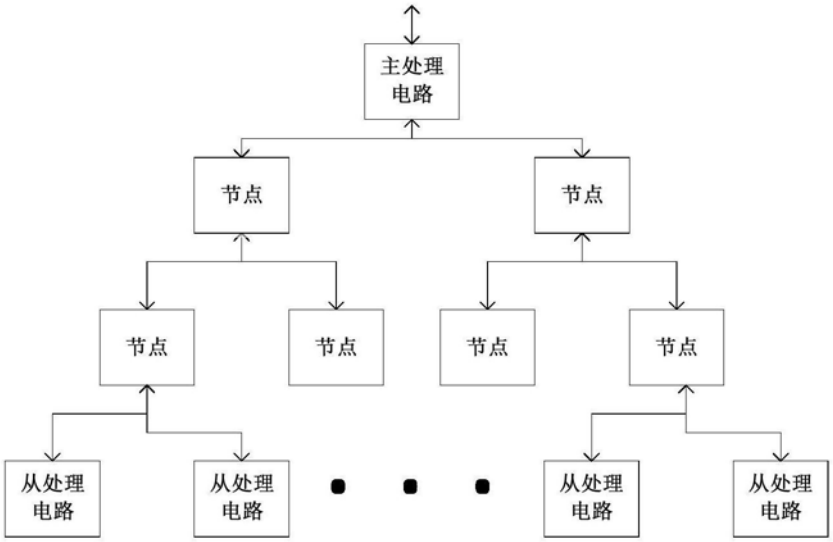

主处理电路
节点
节点
节点
节点
节点
节点
从处理电路
从处理电路
从处理电路
从处理电路

图9

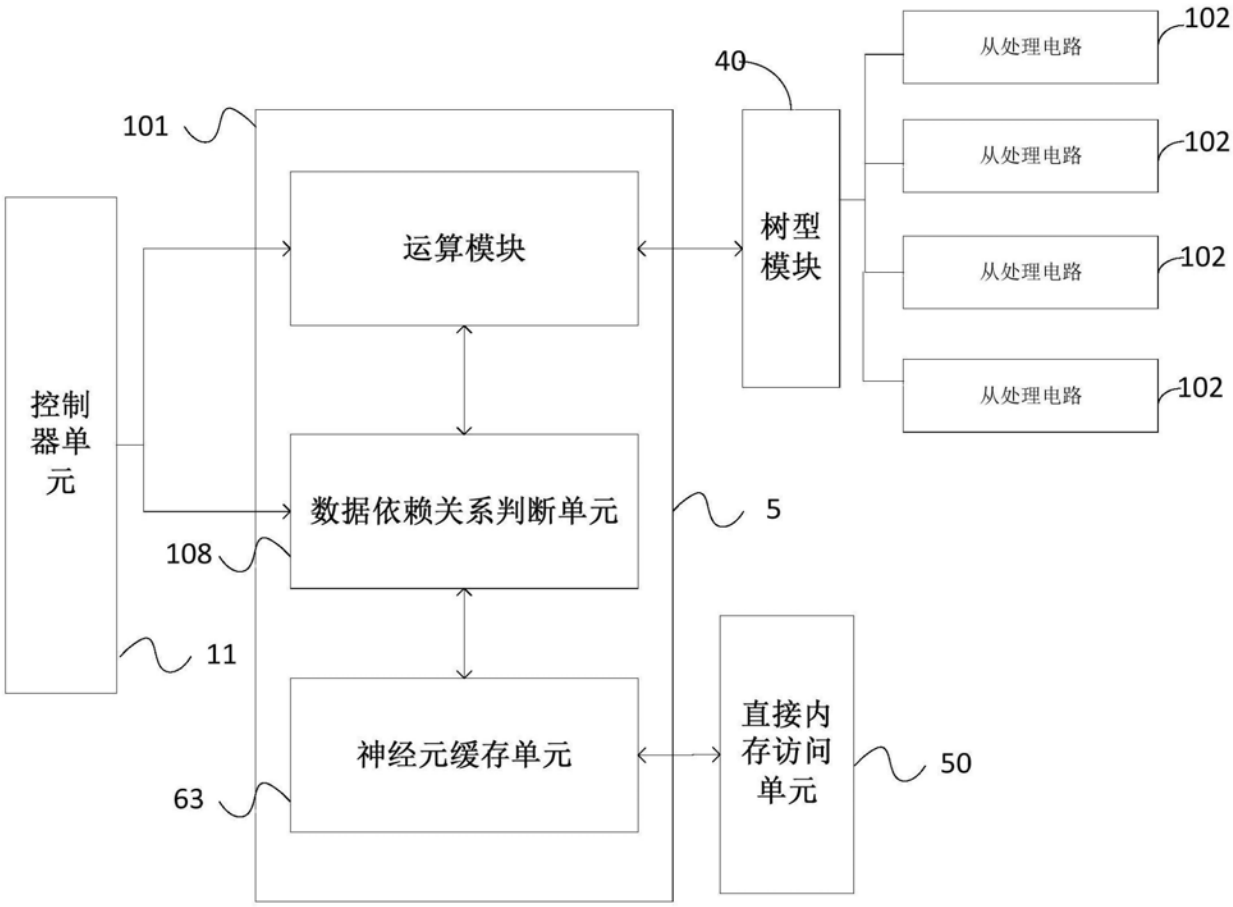

101
40
102
102
102
102
从处理电路
从处理电路
从处理电路
从处理电路
运算模块
树型模块
控制器单元
数据依赖关系判断单元
5
108
11
神经元缓存单元
直接内存访问单元
50
63

图10

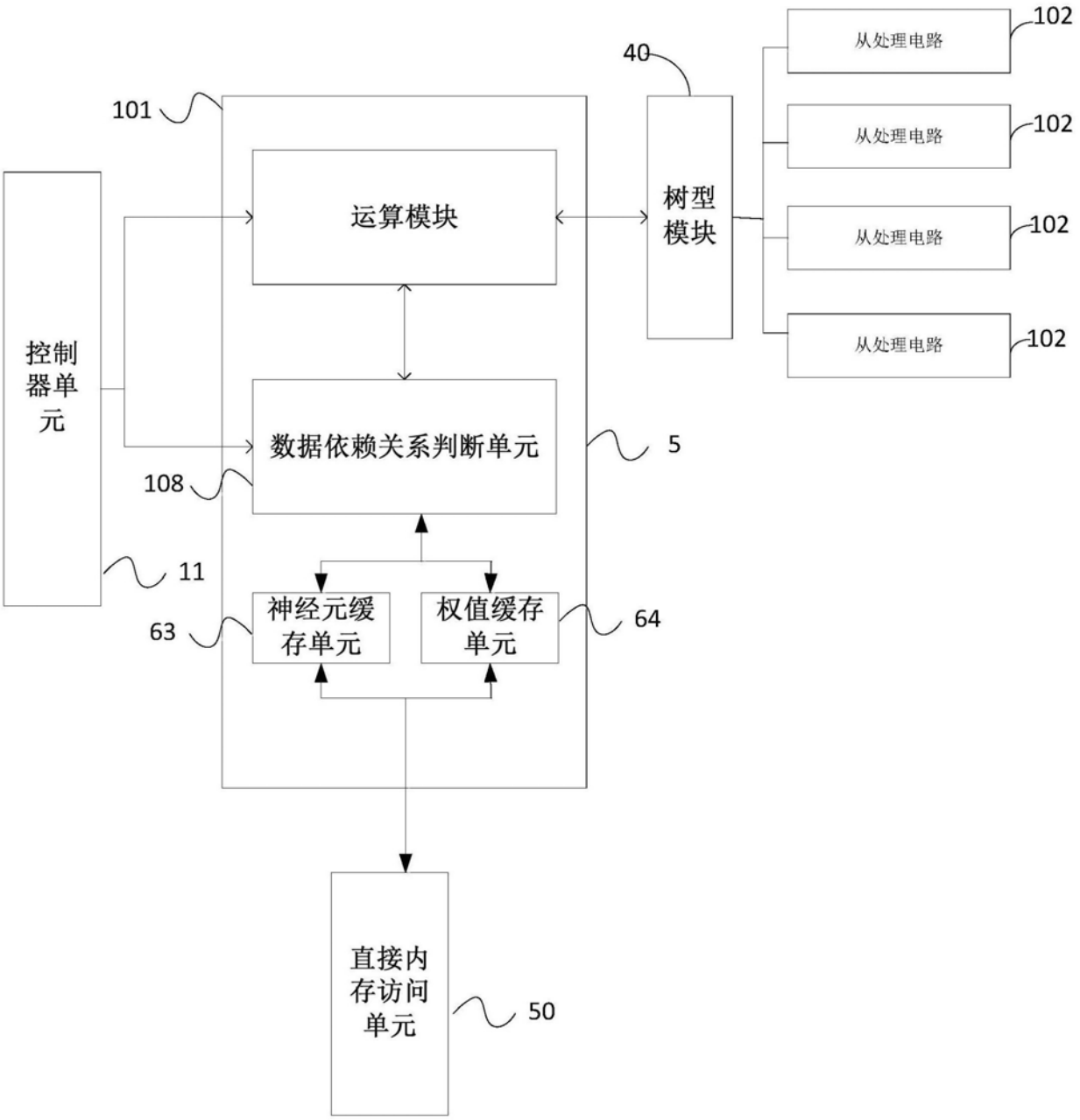

101
运算模块
控制器单元
数据依赖关系判断单元
108
5
11
神经元缓存单元
63
权值缓存单元
64
直接内存访问单元
50
40
树型模块
从处理电路
102
从处理电路
102
从处理电路
102
从处理电路
102

图11

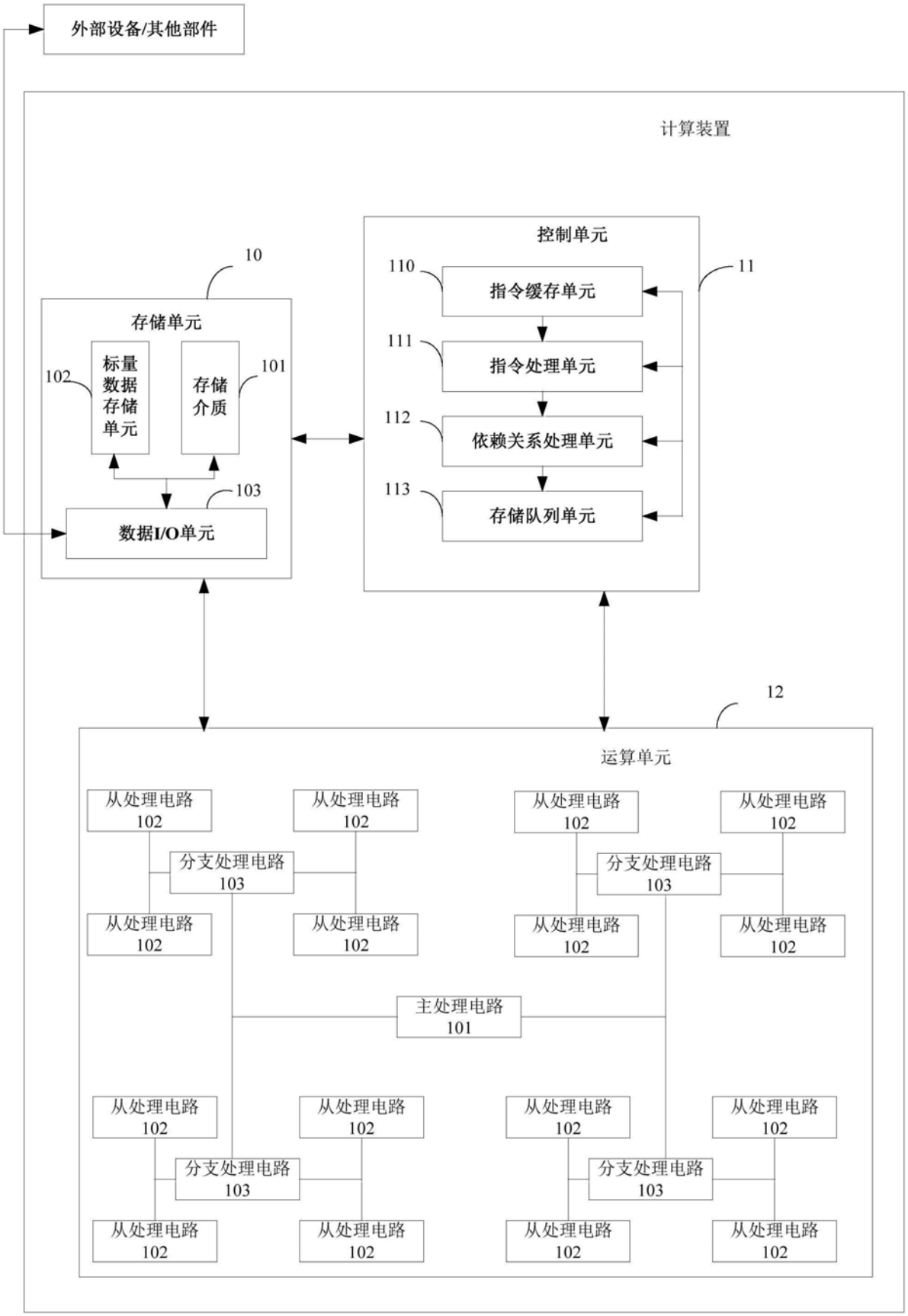

外部设备/其他部件
计算装置
控制单元
10
11
110
指令缓存单元
111
指令处理单元
112
依赖关系处理单元
113
存储队列单元
存储单元
102
标量数据存储单元
存储介质
101
103
数据I/O单元
12
运算单元
从处理电路 102
分支处理电路 103
主处理电路 101

图12

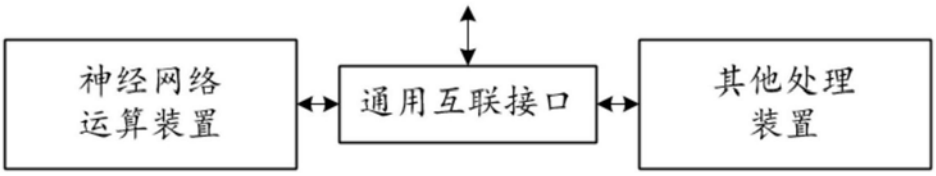


图13

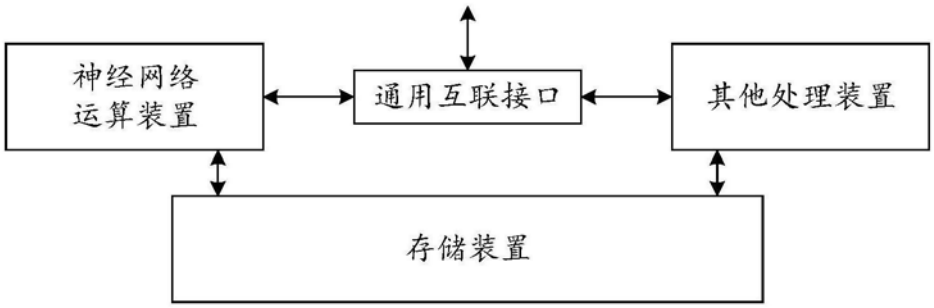


图14

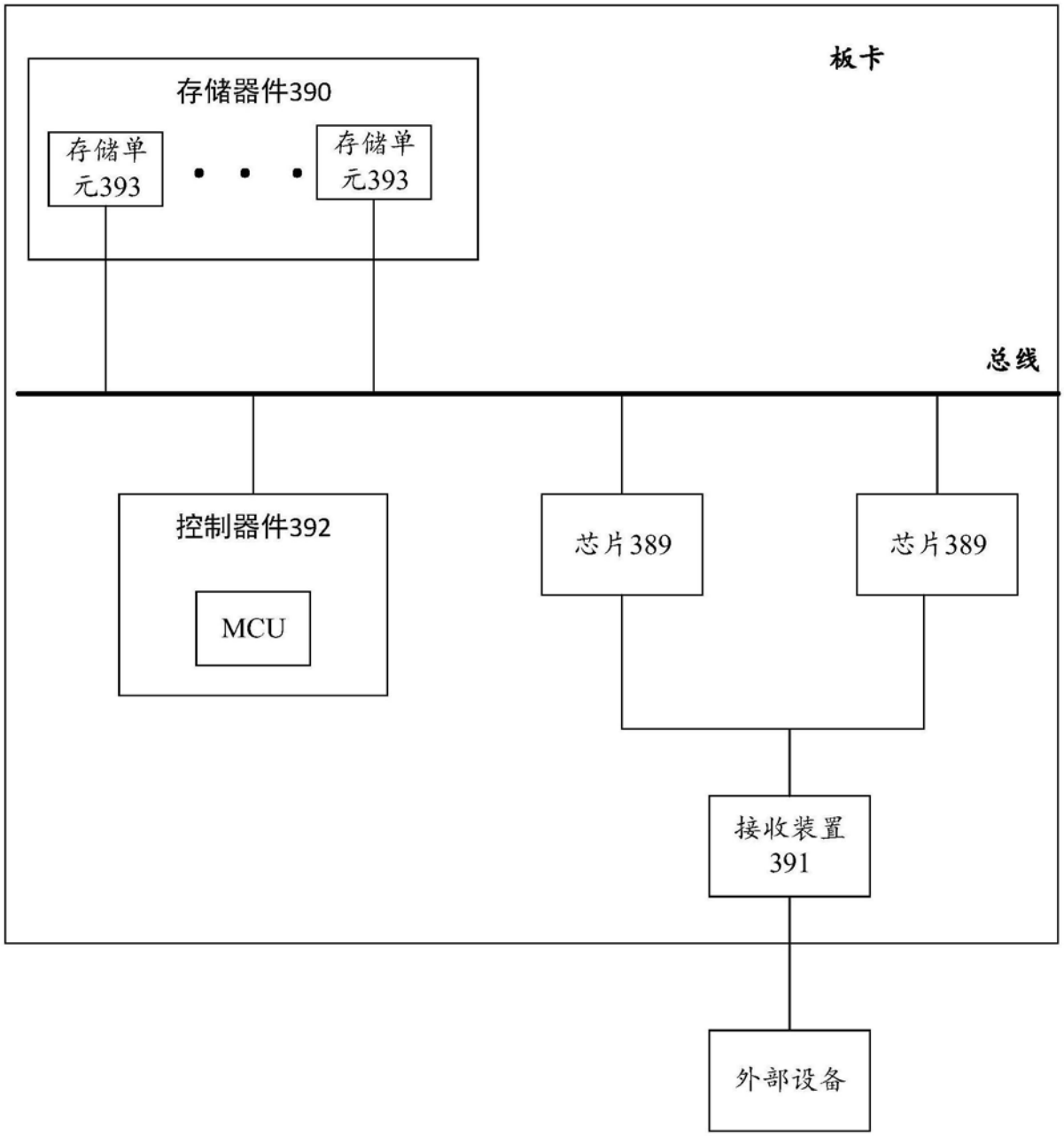

板卡
存储器件390
存储单元393
存储单元393
总线
控制器件392
MCU
芯片389
芯片389
接收装置391
外部设备

图15

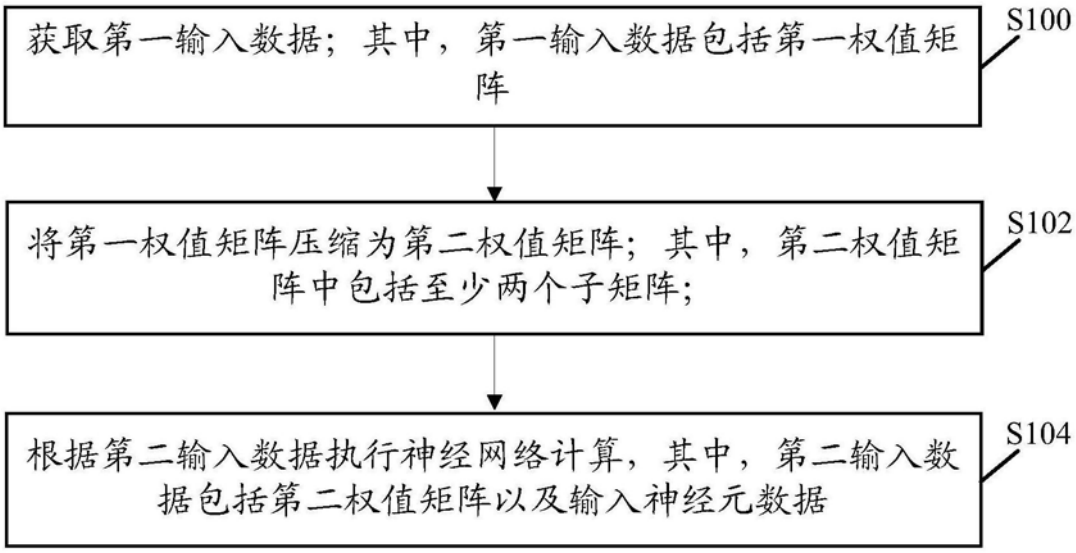

获取第一输入数据；其中，第一输入数据包括第一权值矩阵
S100
将第一权值矩阵压缩为第二权值矩阵；其中，第二权值矩阵中包括至少两个子矩阵；
S102
根据第二输入数据执行神经网络计算，其中，第二输入数据包括第二权值矩阵以及输入神经元数据
S104

图16

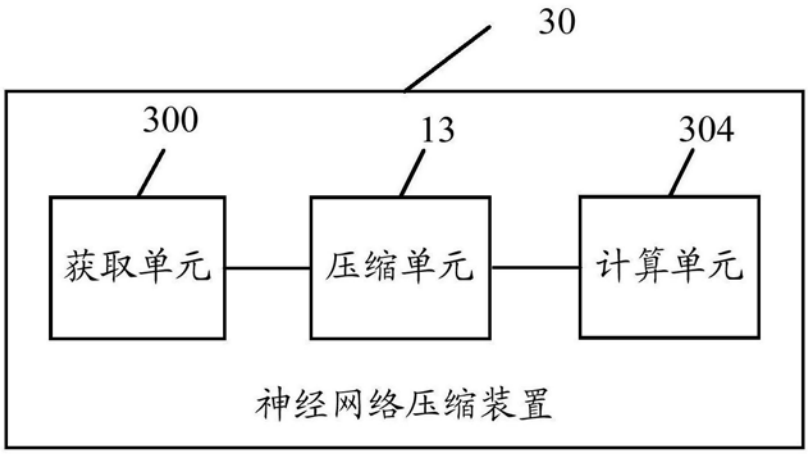

30
300
13
304
获取单元
压缩单元
计算单元
神经网络压缩装置

图17A

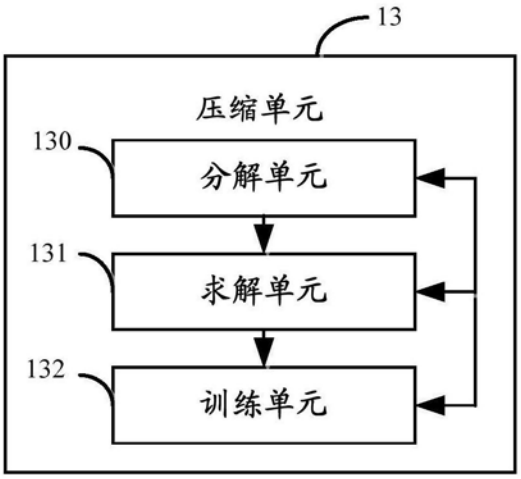

13
压缩单元
130
分解单元
131
求解单元
132
训练单元

图17B

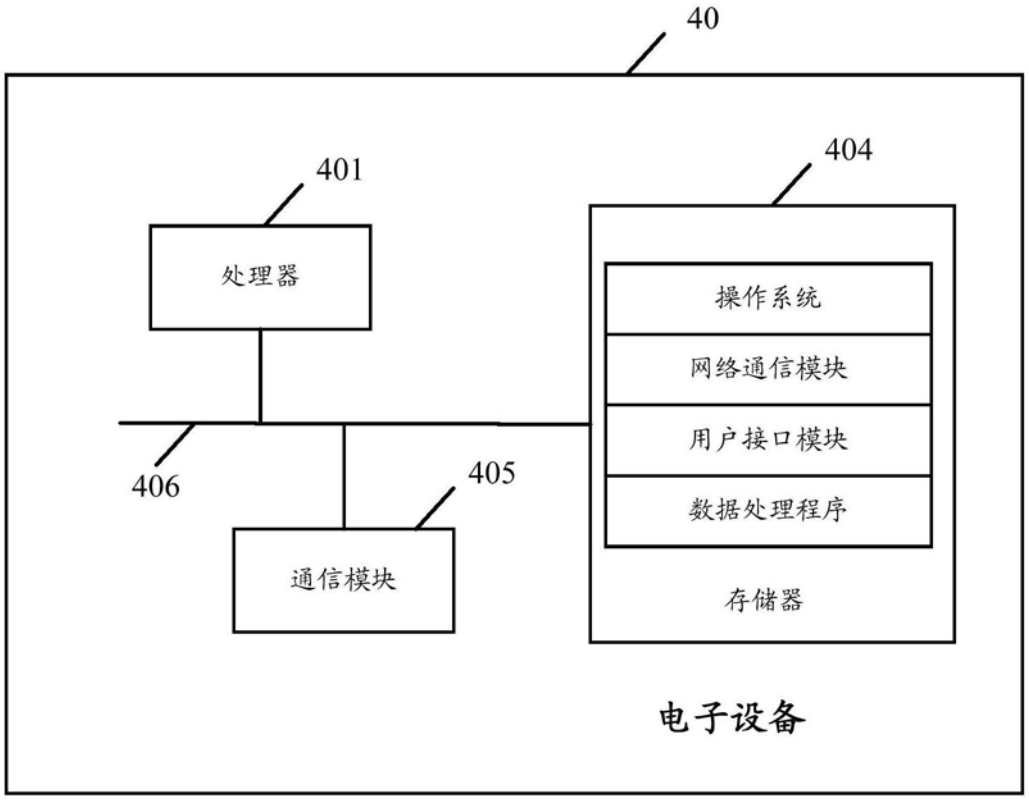

40
401
处理器
404
操作系统
网络通信模块
用户接口模块
数据处理程序
存储器
406
405
通信模块
电子设备

图18