(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 117151177 A
(43)申请公布日 2023.12.01

(21)申请号 202311093908.7

(22)申请日 2018.09.07

(30)优先权数据
62/556,312 2017.09.08 US
16/124,104 2018.09.06 US

(62)分案原申请数据
201880055138.8 2018.09.07

(71)申请人 罗希特·塞思
地址 加拿大安大略省

(72)发明人 罗希特·塞思

(74)专利代理机构 中原信达知识产权代理有限责任公司 11219
专利代理师 达小丽 夏凯

(51)Int.Cl.
G06N 3/063(2023.01)
G06N 3/045(2023.01)
G06N 3/088(2023.01)
G06V 10/764(2022.01)
G06V 10/82(2022.01)
G06F 18/243(2023.01)

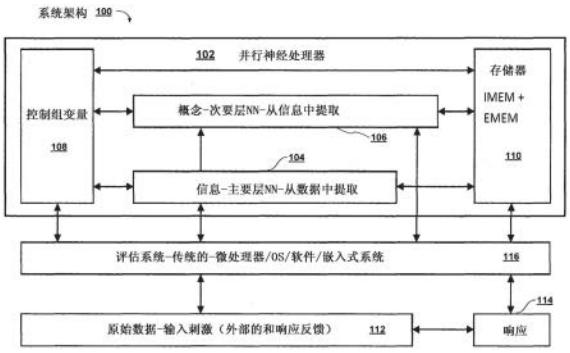
权利要求书5页 说明书16页 附图16页

(54)发明名称

用于人工智能的并行神经处理器

(57)摘要

本发明提供一种用于人工智能的并行神经处理器。本文提供了用于实现专门设计用于并行AI处理的人工神经网络的高效和直观的方法的系统和/或设备。在各种实施方式中,所公开的系统、装置和方法补充或替代用于并行神经处理的传统的系统、设备和方法,该并行神经处理(a)大大地减少处理更复杂问题集合所需的神经处理时间;(b)实现自学习所需的神经可塑性;(c)引入注入直觉元素所需要的除了显式存储器之外的隐式存储器的概念和应用。利用这些特性,公开的发明的实施方式使得模拟人类意识或认知成为可能。



CN 117151177 A

1.一种被配置成处理输入信号的第一子分类器,所述第一子分类器包括:

加权输入电路,所述加权输入电路被配置成向所述输入信号施加权重以生成加权输入信号;

比较电路,所述比较电路被耦合到所述加权输入电路,所述比较电路被配置成:

在比较电路输入线处接收所述加权输入信号;以及

在比较电路输出线处生成第一输出信号;所述比较电路被进一步配置成:

确定是否所述加权输入信号具有在下窗口范围值和上窗口范围值之间的值;

响应于确定所述加权输入信号具有在所述下窗口范围值和所述上窗口范围值之间的值,在所述比较电路输出线处,将所述第一输出信号设置成具有第一值;以及

响应于确定所述加权输入信号具有不在所述下窗口范围值和所述上窗口范围值之间的值,在所述比较电路输出线处将所述第一输出信号设置成具有与所述第一值不同的第二值。

2.根据权利要求1所述的第一子分类器,其中,所述比较电路包括至少一个运算放大器,所述至少一个运算放大器被配置成接收所述加权输入信号并且设置所述第一输出信号。

3.根据权利要求1所述的第一子分类器,其中,施加到所述输入信号以生成所述加权输入信号的权重基于来自第二子分类器的第二输出信号。

4.根据权利要求1所述的第一子分类器,其中,来自所述第一子分类器的所述第一输出信号被发送到第二子分类器。

5.根据权利要求1所述的第一子分类器,其中,所述加权输入电路被配置成接收控制组信号,并且对所述输入信号施加权重以基于所述控制组信号生成所述加权输入信号。

6.根据权利要求5所述的第一子分类器,其中,所述控制组信号控制包括所述第一子分类器的多个子分类器对所述输入信号的响应度。

7.根据权利要求5所述的第一子分类器,其中,所述控制组信号在操作期间被持续和/或间歇地施加到所述比较电路。

8.根据权利要求5所述的第一子分类器,其中,所述加权输入电路包括:

可变电阻器或可变电流或电压调节器,所述可变电阻器或可变电流或电压调节器被配置成接收所述控制组信号并且基于所述控制组信号来调整所述加权输入信号或影响所述比较电路的刺激灵敏度。

9.根据权利要求1所述的第一子分类器,进一步包括:

存储器电路,所述存储器电路被配置成接收和存储来自所述比较电路的所述第一输出信号,并且向第二子分类器提供所述第一输出信号。

10.根据权利要求1所述的第一子分类器,其中,包括所述第一子分类器的多个子分类器形成自组织映射(SOM)。

11.一种被配置成在一个或多个时钟周期期间处理一个或多个输入信号的分类器系统,包括:

多个子分类器,所述多个子分类器中的每个包括:

加权输入电路,所述加权输入电路被配置成向用于各个时钟周期的各个输入信号施加权重以生成加权输入信号;

比较电路,所述比较电路被耦合到所述加权输入电路,所述比较电路被配置成:

在比较电路输入线处接收所述加权输入信号;以及

在比较电路输出线处生成输出信号;所述比较电路被进一步配置成:

确定是否所述加权输入信号具有在下窗口范围值和上窗口范围值之间的值;

响应于确定所述加权输入信号具有在所述下窗口范围值和所述上窗口范围值之间的值,在所述比较电路输出线处将第一输出信号设置成具有大于预定输出阈值的值;以及

响应于确定所述加权输入信号具有不在所述下窗口范围值和所述上窗口范围值之间的值,在所述比较电路输出线处将所述第一输出信号设置成具有小于所述预定输出阈值的值;

存储器电路,所述存储器电路被配置成接收和缓冲由所述多个子分类器中的每个子分类器的各个比较电路所生成的输出信号;以及

主分类器,所述主分类器被耦合到所述存储器电路,所述主分类器被配置成:

在所述一个或多个时钟周期期间从所述存储器电路接收来自由所述存储器电路所缓冲的所述多个子分类器中的每个子分类器的各个输出信号中的每个,以及

基于产生具有大于所述预定输出阈值的值的各个输出信号的所述多个子分类器的子集来确定分类器响应。

12.根据权利要求11所述的分类器系统,其中,所述子分类器中的每个具有在所述下窗口范围值和所述上窗口范围值之间、不与所述多个子分类器中的任何其他子分类器的任何其他各个窗口范围重叠的各个窗口范围。

13.根据权利要求11所述的分类系统,进一步包括:

复用器,所述复用器被耦合到所述多个子分类器,所述复用器被配置成在单个时钟周期期间向所述多个子分类器提供所述输入信号中的一个。

14.根据权利要求11所述的分类系统,其中,所述加权输入电路被配置为接收控制组信号并且对各个输入信号施加权重以基于控制组信号生成加权输入信号。

15.根据权利要求14所述的分类器系统,其中,所述控制组信号控制各个子分类器对所述各个输入信号的响应度。

16.根据权利要求14所述的分类器系统,其中,所述控制组信号在操作期间被持续和/或间歇地施加到所述比较电路。

17.根据权利要求14所述的分类系统,其中,所述加权输入电路包括:

可变电阻器或可变电流或电压调节器,所述可变电阻器或可变电流或电压调节器被配置成接收所述控制组信号并且基于所述控制组信号来调整所述加权输入信号或影响所述比较电路的刺激灵敏度。

18.根据权利要求11所述的分类器系统,其中,所述多个子分类器形成自组织映射(SOM)。

19.一种使用第一子分类器处理输入信号的方法,所述第一子分类器包括加权输入电路和被耦合到所述加权输入电路的比较电路,所述方法包括:

在所述加权输入电路处向所述输入信号施加权重以生成加权输入信号;

在所述比较电路处,在比较电路输入线处接收所述加权输入信号;

在所述比较电路处,在比较电路输出线处生成第一输出信号,包括:

在所述比较电路处,确定是否所述加权输入信号具有在下窗口范围值和上窗口范围值之间的值;

响应于确定所述加权输入信号具有在所述下窗口范围值和所述上窗口范围值之间的值,在所述比较电路处,在所述比较电路输出线处将所述第一输出信号设置成具有第一值;以及

响应于确定所述加权输入信号具有不在所述下窗口范围值和所述上窗口范围值之间的值,在所述比较电路处,在所述比较电路输出线处将所述第一输出信号设置为具有不同于所述第一值的第二值。

20.根据权利要求19所述的方法,进一步包括:

在所述加权输入电路处接收控制组信号;以及

在所述加权输入电路处,向所述输入信号施加权重以基于所述控制组信号生成所述加权输入信号。

21.根据权利要求20所述的方法,其中,所述控制组信号控制包括所述第一子分类器的多个子分类器对所述输入信号的响应度。

22.根据权利要求20所述的方法,其中,所述控制组信号在操作期间被持续和/或间歇地施加到所述比较电路。

23.根据权利要求20所述的方法,其中,所述加权输入电路包括可变电阻器或可变电流或电压调节器,所述方法进一步包括:

在所述可变电阻器或可变电流或电压调节器处:

接收所述控制组信号,并且基于所述控制组信号来调整所述加权输入信号或影响所述比较电路的刺激灵敏度。

24.根据权利要求19所述的方法,其中,所述比较电路包括至少一个运算放大器电路,所述方法进一步包括:

在所述至少一个运算放大器电路处接收所述加权输入信号,并且在所述至少一个运算放大器电路处,设置所述第一输出信号。

25.根据权利要求19所述的方法,进一步包括:

在所述加权输入电路处接收来自第二子分类器的第二输出信号。

26.根据权利要求19所述的方法,进一步包括:

从所述第一子分类器向第二子分类器发送所述第一输出信号。

27.根据权利要求19所述的方法,其中,所述加权输入电路包括可变电阻器电路,所述方法进一步包括:

在可变电阻器电路处接收所述控制组信号,并且基于所述控制组信号来调整所述加权输入信号。

28.根据权利要求19所述的方法,其中,所述第一子分类器包括被耦合到所述比较电路的存储器电路,所述方法进一步包括:

在所述存储器电路处接收和存储来自所述比较电路的所述第一输出信号,并且向第二子分类器提供所述第一输出信号。

29.根据权利要求19所述的方法,其中,包括所述第一子分类器的多个子分类器形成自组织映射(SOM)。

30.一种使用分类器系统在一个或多个时钟周期期间处理一个或多个输入信号的方法,所述分类器系统包括多个子分类器和被耦合到所述多个子分类器的主分类器,所述多个子分类器均包括加权输入电路和比较电路,所述方法包括:

在每个子分类器处:

在所述加权输入电路处,向用于各个时钟周期的各个输入信号施加权重以生成加权输入信号;

在所述比较电路处,在比较电路输入线处接收所述加权输入信号;以及

在所述比较电路处,在比较电路输出线处生成输出信号,包括:

在所述比较电路处,确定是否所述加权输入信号具有在下窗口范围值和上窗口范围值之间的值;

响应于确定所述加权输入信号具有在所述下窗口范围值和所述上窗口范围值之间的值,在所述比较电路处,在所述比较电路输出线处将输出信号设置成具有大于预定输出阈值的值;以及

响应于确定所述加权输入信号具有不在所述下窗口范围值和所述上窗口范围值之间的值,在所述比较电路处,在所述比较电路输出线处将所述输出信号设置成具有小于所述预定输出阈值的值;以及

在所述主分类器处:

在所述一个或多个时钟周期期间从所述多个子分类器中的每个接收所述输出信号中的每个,以及

基于产生具有大于所述预定输出阈值的值的各个输出信号的所述多个子分类器的子集来确定分类器响应。

31.根据权利要求30所述的方法,其中,所述子分类器中的每个具有在所述下窗口范围值和所述上窗口范围值之间、不与所述多个子分类器中的任何其他子分类器的任何其他各个窗口范围重叠的各个窗口范围。

32.根据权利要求30所述的方法,其中,所述分类器系统包括被耦合到所述多个子分类器的复用器电路,进一步包括:

在所述复用器电路处,在单个时钟周期期间向所述多个子分类器提供所述输入信号中的一个。

33.根据权利要求30所述的方法,进一步包括:

在每个子分类器处:

在所述加权输入电路处接收控制组信号;

在所述加权输入电路处,向用于各个时钟周期的各个输入信号施加权重以基于所述控制组信号生成所述加权输入信号。

34.根据权利要求33所述的方法,其中,所述控制组信号控制各个子分类器对所述各个输入信号的响应度。

35.根据权利要求33所述的方法,其中,所述控制组信号在操作期间被持续和/或间歇地施加到所述比较电路。

36.根据权利要求33所述的方法,其中,所述加权输入电路包括可变电阻器或可变电流或电压调节器,所述方法进一步包括:

在所述可变电阻器或可变电流或电压调节器处：

接收所述控制组信号，并且基于所述控制组信号来调整所述加权输入信号或影响所述比较电路的刺激灵敏度。

37.根据权利要求30所述的方法，其中，所述多个子分类器形成自组织映射(SOM)。

用于人工智能的并行神经处理器

[0001] 本申请是2020年2月25日提交的申请号为201880055138.8(PCT/IB2018/000994)、申请日为2018年9月7日、标题为“用于人工智能的并行神经处理器”的专利申请的分案申请。

技术领域

[0002] 所公开的实施方式通常涉及人工智能,并且更具体地涉及一种用于实现用于人工智能的并行神经处理器的方法、系统和设备。

背景技术

[0003] 人工智能(AI)应用传统上被设计用于和作为软件驱动系统。在这样的系统中,处理元件(在AI“脑”中起到“神经元”的作用)被编程以占据硬件存储器中的固定状态。神经元通过耦合变量的代表值互连,以形成人工神经网络。这些神经元的状态是使用处理权重、偏差和输入数据的激活函数迭代评估的,以产生二进制输出值(即,0或1)。神经元的合成状态作为神经元的输出状态被存储在存储器中,随后用作人工神经网络中连接的神经元的输入。在高水平处,以迭代的方式评估多个神经元的输入和输出状态。一些系统使用多线程和多核处理器以同时评估神经元的多个块,但系统作为整体本质上保持“串行”的。较大的神经网络比较小的网络能够解决更复杂的和各种各样的问题。但较大的神经网络需要具有较大的核心数量和/或较大的线程数量的微处理器。因此,AI受这些传统处理器的速度的限制。

[0004] 为了弥补传统硬件的限制,精心设计具有各种快捷方式和边界条件的AI系统,并针对特定的问题集合进行了调整。由于边界条件是预先限定的,这些系统限于高度特定的应用。例如,被训练用于识别人脸的AI系统,可能不能有效地识别长颈鹿的脸。

[0005] 传统的系统产生大的开销,不能以节省成本的方式实现快速响应复杂的问题集合,而且远远不能实现人工意识。

发明内容

[0006] 因此,需要用于实现特别设计用于并行AI处理的人工神经网络的更高效和直观的方法的系统和/或设备。在一些实施方式中,所公开的系统、设备和方法补充或替代用于并行神经处理的传统的系统、设备和方法,该并行神经处理(a)大大地减少处理更复杂的问题集合所需的神经处理时间;(b)实现自学习所需的神经可塑性;以及(c)引入注入直觉元素所需要的除了显式存储器之外的隐式存储器的概念和应用。利用这些特性,公开的本发明的一些实施方式使得模拟人类意识或认知成为可能。

[0007] (A1)在一个方面中,一些实施方式包括被配置成处理输入信号的第一子分类器。第一子分类器包括被配置成向输入信号施加权重以生成加权输入信号的加权输入模块。第一子分类器还包括被耦合到加权输入模块的比较模块。比较模块被配置为:在比较模块输入线处接收加权输入信号;以及在比较模块输出线处生成第一输出信号。比较模块进一步

被配置成:确定是否加权输入信号具有在下窗口范围值和上窗口范围值之间的值。响应于确定加权输入信号具有在下窗口范围值和上窗口范围值之间的值,比较模块被配置成:在比较模块输出线处将第一输出信号设置成具有第一值。响应于确定加权输入信号具有不在下窗口范围值和上窗口范围值之间的值,比较模块被配置成:在比较模块输出线处将第一输出信号设置成具有与第一值不同的第二值。

[0008] (A2)在A1的第一子分类器的一些实施方式中,比较模块包括被配置成接收加权输入信号和设置第一输出信号的至少一个运算放大器。

[0009] (A3)在A1的第一子分类器的一些实施方式中,施加于输入信号以生成加权输入信号的权重基于来自第二子分类器的第二输出信号。

[0010] (A4)在A1的第一子分类器的一些实施方式中,来自第一子分类器的第一输出信号被发送到第二子分类器。

[0011] (A5)在A4的第一子分类器的一些实施方式中,加权输入模块被配置成接收控制组信号并且对输入信号施加权重以基于控制组信号生成加权输入信号。

[0012] (A6)在A5的第一子分类器的一些实施方式中,加权输入模块包括被配置成接收控制组信号并且基于控制组信号调整加权输入信号的可变电阻器。

[0013] (A7)在A1的第一子分类器的一些实施方式中,第一子分类器进一步包括被配置成接收和存储来自比较模块的第一输出信号并且向第二子分类器提供第一输出信号的存储器模块。

[0014] (A8)在另一方面中,一些实施方式包括被配置成在一个或多个时钟周期期间处理一个或多个输入信号的分类器系统。分类器系统包括多个子分类器。多个子分类器中的每个包括被配置成向用于各个时钟周期的各个输入信号施加权重以生成加权输入信号的加权输入模块。多个子分类器中的每个还包括耦合到加权输入模块的比较模块。比较模块被配置成在比较模块输入线处接收加权输入信号,并且在比较模块输出线处生成输出信号。比较模块进一步被配置成确定是否加权输入信号具有在下窗口范围值和上窗口范围值之间的值。比较模块进一步被配置成响应于确定加权输入信号具有在下窗口范围值和上窗口范围值之间的值,在比较模块输出线处将第一输出信号设置成具有大于预定输出阈值的值。比较模块进一步被配置成响应于确定加权输入信号具有不在下窗口范围值和上窗口范围值之间的值,在比较模块输出线处将第一输出信号设置成具有小于预定输出阈值的值。分类器系统还包括耦合到多个子分类器的主分类器。主分类器被配置成在一个或多个时钟周期期间从多个子分类器中的每个接收各个输出信号中的每个,以及基于产生具有大于预定输出阈值的值的各个输出信号的多个子分类器的子集来确定分类器响应。

[0015] (A9)在(A8)的分类器系统的一些实施方式中,子分类器中的每个具有在下窗口范围值和上窗口范围值之间的不与任何其他子分类器的任何其他各个窗口范围重叠的各个窗口范围。

[0016] (A10)在(A8)的分类器系统的一些实施方式中,分类器系统进一步包括耦合到多个子分类器的复用器,该复用器被配置为在单个时钟周期期间向多个子分类器提供输入信号中的一个。

[0017] (A11)在另一方面中,一些实施方式包括使用第一子分类器处理输入信号的方法。第一子分类器包括加权输入模块和耦合到加权输入模块的比较模块。该方法包括在加权输

入模块处向输入信号施加权重以生成加权输入信号。该方法进一步包括在比较模块处,在比较模块输入线处接收加权输入信号。该方法进一步包括在比较模块通过电处理在比较模块输出线处生成第一输出信号。电处理可以被国家化(nationalized)为步骤序列,该步骤序列包括在比较模块处确定是否加权输入信号具有在下窗口范围值和上窗口范围值之间的值。该步骤序列进一步包括响应于确定加权输入信号具有在下窗口范围值和上窗口范围值之间的值,在比较模块处,在比较模块输出线处将第一输出信号设置成具有第一值。该步骤序列进一步包括响应于确定加权输入信号具有不在下窗口范围值和上窗口范围值之间的值,在比较模块处,在比较模块输出线处将第一输出信号设置成具有与第一值不同的第二值。

[0018] (A12)在(A11)的方法的一些实施方式中,比较模块包括至少一个运算放大器模块,并且该方法进一步包括在至少一个运算放大器模块处接收加权输入信号和在至少一个运算放大器模块处设置第一输出信号。

[0019] (A13)在(A11)的方法的一些实施方式中,该方法进一步包括在加权输入模块处从第二子分类器接收第二输出信号。

[0020] (A14)在(A11)的方法的一些实施方式中,该方法进一步包括将第一输出信号从第一子分类器发送到第二子分类器。

[0021] (A15)在(A14)的方法的一些实施方式中,该方法进一步包括在加权输入模块处接收控制组信号,并且在加权输入模块处将权重施加到输入信号以基于控制组信号生成加权输入信号。

[0022] (A16)在(A15)的方法的一些实施方式中,加权输入模块包括电流或电压控制器(可变电阻器模块、电阻器梯(resistor ladder)、电阻器网络或用于控制电流的电路),并且该方法进一步包括在可变电阻器模块处接收控制组信号并且基于控制组信号来调整加权输入信号。

[0023] (A17)在(A13)的方法的一些实施方式中,第一子分类器包括耦合到比较模块的存储器模块,并且该方法进一步包括在存储器模块处接收和存储来自比较模块的第一输出信号,并且向第二子分类器提供第一输出信号。

[0024] (A18)然而在另一方面中,一些实施方式包括在一个或多个时钟周期期间使用分类器系统处理一个或多个输入信号的方法。分类器系统包括多个子分类器和耦合到多个子分类器的主分类器,多个子分类器均包括加权输入模块和比较模块。该方法包括在每个子分类器处,在加权输入模块处,向用于各个时钟周期的各个输入信号施加权重以生成加权输入信号;在比较模块处,通过在比较模块输入线处接收加权输入信号;以及在比较模块处,通过电处理在比较模块输出线处生成输出信号。该处理可以被国家化为步骤序列,该步骤序列包括在比较模块处,确定是否加权输入信号具有在下窗口范围值和上窗口范围值之间的值。该步骤序列进一步包括响应于确定加权输入信号具有在下窗口范围值和上窗口范围值之间的值,在比较模块处,在比较模块输出线处将输出信号设置成具有大于预定输出阈值的值。该步骤序列进一步包括响应于确定加权输入信号具有不在下窗口范围值和上窗口范围值之间的值,在比较模块处,在比较模块输出线处将输出信号设置成具有小于预定输出阈值的值。该方法进一步包括在主分类器处:在一个或多个时钟周期期间接收来自多个子分类器中的每个的输出信号,以及基于产生具有大于预定输出阈值的值的各个输出信

号的多个子分类器的子集来确定分类器响应。

[0025] (A19)在(A18)的方法的一些实施方式中,子分类器中的每个具有在下窗口范围值和上窗口范围值之间、不与任何其他子分类器的任何其他各个窗口范围重叠的各个窗口范围。

[0026] (A20)在(A18)的方法的一些实施方式中,分类器系统包括耦合到多个子分类器的复用器模块,并且该方法进一步包括在复用器模块处在单个时钟周期期间向多个子分类器提供输入信号中的一个。

附图说明

[0027] 为了更好地理解所描述的各种实施方式,应结合以下附图参考以下实施方式的描述,并且其中贯穿附图相同的附图标记指相应的部分。

[0028] 图1A和1B是示出了根据一些实施方式的具有并行神经处理(PNP)AI处理器的示例系统架构的框图。

[0029] 图2A示出了根据一些实施方式的模拟窗口比较器,PNP AI处理器的组件;以及图2B示出了根据一些实施方式的非反相窗口比较器。

[0030] 图3示出了根据一些实施方式的神经网络中的一系列窗口比较器电路。

[0031] 图4示出了根据一些实施方式的在具有数据流控制阶段的互连神经网络中的一系列窗口比较器。

[0032] 图5示出了根据一些实施方式的具有用于控制神经网络的神经可塑性和行为的添加的控制组(CG)的图4中的窗口比较器系列。

[0033] 图6示出了根据一些实施方式的具有通过CG和神经网络可寻址的添加的隐式存储器的图5中的窗口比较器系列。

[0034] 图7是示出了根据一些实施方式的具有图1A或1B的并行神经处理器的代表性系统700的框图。

[0035] 图8A-8D示出了根据一些实施方式的使用包括加权输入模块和耦合到加权输入模块的比较模块的子分类器来处理输入信号的方法的流程图表示。

[0036] 图9A-9E示出了根据一些实施方式的使用分类器在一个或多个时钟周期期间处理一个或多个输入信号的方法的流程图,该分类器包括多个子分类器和耦合到多个子分类器的主分类器,多个子分类器均包括加权输入模块和比较模块。

具体实施方式

[0037] 现在将详细参考实施方式,其示例在附图中示出。在以下的详细描述中,阐述了许多具体细节,以便提供对所描述的各种实施方式的透彻理解。然而,对于本领域的普通技术人员来说,显而易见的是,在没有这些具体细节的情况下,可以实践各种所描述的实施方式。在其他实例中,没有详细描述众所周知的方法、程序、组件、电路和网络,以免不必要地模糊实施方式的各个方面。

[0038] 还应理解,尽管在某些实例中,本文使用的术语第一、第二等用于描述各种元素,但这些元素不应受这些术语的限制。这些术语仅用于区分一个元素和另一个元素。例如,在不脱离所描述的各种实施方式的范围的情况下,第一电子设备可以被称为第二电子设备,

并且,类似地,第二电子设备可以被称为第一电子设备。第一电子设备和第二电子设备都是电子设备,但它们不必然是相同电子装置。

[0039] 在本文所描述的各种实施方式的描述中使用的术语仅用于描述特定实施方式的目的,并不旨在限制。如在对各种所述实施方式和所附权利要求的描述中使用的单数形式“一”、“一个”和“该”意在包括复数形式,除非上下文另有明确指示。还应理解,本文中使用的术语“和/或”是指并涵盖一个或多个相关的所列项目的任何一个和所有可能的组合。应进一步理解,当在本说明书中使用时,术语“包括(include)”、“包括(including)”、“包含(comprise)”和/或“包含(comprising)”时,指定所述特征、整数、步骤、操作、元素和/或组件的存在,但不排除一个或多个其他特征、整数、步骤、操作、元件、组件和/或其组的存在或添加。

[0040] 如本文所用,取决于上下文,术语“如果”可选地被解释为指“当”或“此后”或“响应于确定”或“响应于检测”或“根据确定”。类似地,取决于上下文,短语“如果已确定”或“如果[陈述的条件或事件]被检测到”可选地被解释为指“在确定之后”或“响应于确定”或“在检测到[陈述的条件或事件]之后”或“响应于检测到[陈述的条件或事件]”或“根据检测到[陈述的条件或事件]的确定”

[0041] 图1A和1B是根据一些实施方式示出具有并行神经处理(PNP)AI处理器的示例系统架构的框图。图1A是示出根据一些实施方式的将PNP AI处理器102与处理原始数据、向PNP AI处理器102发送AI任务(例如,图像识别、自然语言处理)以及评估PNP AI处理器102的输出的评估系统116集成的示例系统架构100的框图。评估系统116将任务发送到PNP AI处理器102,允许系统对有意义的响应114的外部刺激112快速反应。在该配置中,PNP AI处理器102用作执行从评估系统116接收到的关键任务的协处理器。在一些实施方式中,评估系统116包括传统的微处理器、软件、嵌入式或移动应用。图1B是示出示例系统架构120的框图,其中PNP AI处理器102响应于输入刺激112而独立操作(没有图1A的评估系统116的帮助),处理输入刺激112并生成响应114。在该配置中,如虚线所指示的,PNP AI处理器102还可以与包括现存的微处理器、软件、嵌入式或移动应用的传统系统一起操作。

[0042] 当PNP AI处理器102充当协处理器时,如图1A所示,根据一些实施方式,评估系统116(本文中有时称为处理器子系统)负责预处理人来自外部输入设备的原始数据输入(本文中有时称为刺激),并将原始数据输入发送到PNP AI处理器102。例如,处理器子系统将输入数据转换成输入到PNP AI处理器的代表性电压电平。在各个实施方式中,这样的刺激包括由用户输入(例如,键盘或鼠标)的数据、来自外部设备(例如,照相机)的图像或视频输入、音频输入、来自传感器和/或马达的传感数据。此列表是示例性的,并非详尽的。作为例证,当由系统架构100控制的可移动的机械臂由于外部的物理约束(例如,被墙阻挡)而卡住时,机械臂中的一个或多个传感器可以向系统生成影响控制组(CG)模块108(以下参照图6详细描述)的反馈。

[0043] 如图1A所示的评估系统116,根据一些实施方式,结合PNP AI处理器实现交互式智能系统。根据一些实施方式,评估系统根据存储的经验的集合响应于刺激来评估从PNP AI处理器102(本文中有时称为分类器或分类器系统)接收的动作选项。在一些实施方式中,评估系统116基于存储的经验的集合对动作选项的成功概率进行分类,并返回评级最高的动作(即最高的成功概率的动作)。随后,在一些实施方式中,评估系统116生成对应于评级最

高的(刺激的、隐式的、反射的)动作的动作响应(例如,启动移动机械臂的马达)。在一些实施方式中,评估系统116将存储在存储器110的EMEM中的经验数据与存储在存储器110的IMEM中的数据进行比较,并且如果(评估的、显式的、认知的)两个数据之间存在匹配,则生成动作。在一些这样的实施方式中,如果显式和隐式响应不匹配,则评估系统调整(神经可塑性)CG模块108中的一个或多个因素或变量(以下参考图5描述)。如以下进一步详细说明的,来自CG模块108的因素或变量影响在处理时钟周期期间使用的神经元的数量以及神经元的学习速率。此外,在一些实施方式中,如果刺激导致成功的动作,则指示成功的动作的信息被反馈到PNP AI处理器(例如,经由到存储器块110和/或CG模块108的反馈信号),以在随后的处理步骤中进一步使用。在一些实施方式中,对应于刺激所采取的动作导致进一步的刺激。

[0044] 现在注意到PNP AI处理器102。如图1A和1B所示,在一些实施方式中,PNP AI处理器102包含一个或多个神经网络层(例如,从数据输入或刺激中提取信息的神经网络104的信息层或主要层(primary layer),和/或从主要层104的信息输出中提取概念的概念层或次要层(secondary layer))、控制组(CG)模块108和一个或多个存储器块110(例如,隐式存储器块IMEM、显式存储器块EMEM)。在一些实施方式中,CG模块108和/或存储器块110是可选组件。

[0045] 每个层104、106可以包括多个互连的神经元(本文中也称为子分类器)。在一些实施方式中,神经元是基于可配置的拓扑结构连接的。神经元(以下参考图2-6详细说明)是神经网络的处理元件或引擎。与软件神经元在人工神经网络中操作的方式类似,并行神经处理器102中的硬件神经元也对数据进行处理和分类。然而,与软件神经元不同,硬件神经元并行操作,有时一次以数百万、数十亿甚至万亿的集合。在一些实施方式中,层中的神经元网络是以阶段组织的,在给定的时钟周期期间中,其中阶段中的每个神经元集合并行操作。例如,神经元的第一阶段对输入进行操作,接着神经元的一个或多个阶段(有时称为隐藏层)连续地处理神经元的第一阶段的输出,并且最后馈送到神经元的输出阶段。根据一些实施方式,每个阶段需要一个时钟周期来完成,阶段中的所有神经元对相同的输入进行操作。因此,与完全由软件实现的系统相比,这些实施方式实现更快的处理速度。对于输入阶段,根据一些实施方式,对神经处理器来说是外部的或内部的硬件(例如,复用器)跨越神经网络向神经网络供应或分发原始数据。在一些实施方式中,评估系统116预处理原始数据(例如,原始数据112),并向PNP AI预处理器馈送已处理的输入(例如,预定的电压电平)。根据一些实施方式,一旦神经元网络在特定的数据集合上被训练,则不同的神经元组激活以用于不同的数据集合。例如,一个神经元集合(具有第一赢家神经元)响应于接收代表网球图像的输入数据而激活,而另一神经元集合(具有第二赢家神经元)响应于接收代表花朵图像的输入数据而激活。

[0046] 根据一些实施方式,参考以下图5和图6详细说明控制组(CG)模块108。以下参考图6说明存储器块110。根据一些实施方式,PNP AI处理器102的模块之间的各种内部连接和来自PNP AI处理器102的外部连接在图1A中用实心黑线指示。

[0047] 在一些实施方式中,输入刺激112直接连接到并行神经处理器102中的一个或多个神经网络,而不需要评估系统116来预处理原始输入数据。在一些实施方式中,在没有介入中间的评估系统116的情况下,PNP AI处理器102的输出直接连接到响应模块114。在一些实

施方式中,刺激接收模块112和/或响应模块114集成在PNP AI处理器中,即,在没有外部软件/硬件的情况下,PNP AI处理器102可以接收刺激(例如,某种干扰)和/或生成响应(例如,移动机械臂)。

[0048] 这里和以下描述的电路使能应用数百万、数十亿、甚至万亿个神经元的神经网络的并行硬件实施方式。这种大规模并行的硬件实施方式使能解决复杂问题集的人工智能实施方式。

[0049] 在一些实施方式中,神经网络的各个部分以分层的方式组织,以产生在阶段中彼此馈送的多维神经元层。例如,分层神经网络类似于在人类眼睛中神经元网络的组织方式。作为另一示例,神经网络被组织为具有从输入中提取信息的第一神经元数据层(例如,神经元层104,图1A和1B)和基于第一层的输出识别概念的第二神经元数据层(例如,神经元层106,图1A和1B)的两层。在一些这样的实施方式中,具有从低级信息(例如,包含模式的原始数据)中提取和保留高级概念的能力,神经网络处理并生成对新刺激的响应(即,先前在训练时没有看到的数据)。例如,使用上述电路的神经网络可以识别相关的高级概念,诸如“听觉对耳朵如同视觉对眼睛(感官)”可以在较高的连接层中提取,并在诸如味觉的新的、不相关的、没有经验的刺激期间应用。

[0050] 图2A示出了根据一些实施方式的图1A-1B的PNP AI处理器102的模拟窗口比较器(WC)216,其是子分类器的组件。在一些实施方式中,与通过使用微处理器、随机存取存储器、操作系统/嵌入式系统和软件模拟人工神经元的传统神经网络不同,这里公开的系统是硬件实现。根据一些实施方式,诸如图2A中示出的电路的WC电路(本文有时称为比较器)216形成神经元的基础。在一些实施方式中,使用各种基本电子元件(例如,结晶体管、fet)来构造WC电路。在一些实施方式中,使用集成电路(例如,运算放大器)来构造WC电路。如图2A所示,根据一些实施方式,如果施加到WC电路的输入(I_{WC} 206)的电压落在低参考电压(R_L 208)和高参考电压(R_H 210)之间,则WC电路输出(O_{WC} 214)高位(1),并且如果输入(I_{WC} 206)落在高和低参考电压参考电压之外,则输出(/O_{WC} 212)低位(0)。在一些实施方式中,单个WC电路形成如果输入电压在范围内,则激发或激活的单独的“硬件神经元”,产生与经由软件应用实现的虚拟神经元的响应类似的全响应或无响应。

[0051] 图2B示出了根据一些实施方式的非反相WC 218。在一些实施方式中,WC 218执行与图2A中的WC电路216基本相同的功能。在一些实施方式中,WC 218是使用两个连接的运算放大器(图2A的Op-Amp202和Op-Amp204)形成的。

[0052] 在一些实施方式中,WC电路被配置成相对于输入条件反相或非反相。为了简单起见,图2A-2B中仅示出WC的非反相输出(O_{WC} 214)。本文中讨论的示例应视为示例,而不是限制。

[0053] 图3示出了根据一些实施方式的神经网络(例如,图1A-1B的主要层104或次要层106)中的一系列WC电路(WC_1、WC_2、WC_3、…、WC_N)。在一些实施方式中,一系列WC互连以类似于软件神经元形成基于软件的传统神经网络的方式形成神经网络。根据一些实施方式,WC基于神经网络拓扑结构互连。神经网络拓扑表示神经元连接以形成网络的方式。神经网络拓扑结构也可以被看作是借助神经元的连接的神经元之间的关系。WC可以每个在相同的输入刺激I_{WC}(318)上操作以产生相应的输出。

[0054] 在一些实施方式中,每个WC具有窗口电压范围(WVR),其中,如果输入在WVR内,WC

将产生第一值,并且如果输入在WVR外,则产生第二值。在一些实施方式中,WVR是低参考电压和高参考电压之间的差。在一些实施方式中,每个WC具有唯一的WVR。例如,在图3中,比较器WC_1的参考电压(R_L^1 302和R_H^1 304)、比较器WC_2的参考电压(R_L^2 306和R_H^2 308)、比较器WC_3的参考电压(R_L^3 310和R_H^3 312)和比较器WC_N的参考电压(R_L^N 314和R_H^N 316)的参考电压每个被设置使得对应的WVR都是唯一的。在一些实施方式中,WVR是不重叠的。在一些实施方式中,WVR是重叠的,使得多于一个的WC响应于给定的刺激。例如,对于给定的输入刺激I_{WC}(318),多于一个的WC(例如,O_{WC}^1 322、O_{WC}^2 324、O_{WC}^3 326和O_{WC}^N 328)的输出可以等于高值(1)。每个WC的参考电压输入R_H和R_L被加载电压使得$R_H>R_L$,从而为每个WC创建WVR。根据一些实施方式,每个WC在启动时用相应的WVR初始化。在一些实施方式中,每个WC用随机的WVR初始化。在其他一些实施方式中,每个WC用WVR初始化,使得WVRs在整个神经网络上形成均匀的梯度。

[0055] 根据一些实施方式,图4示出了具有数据流控制阶段的互连神经网络(例如,图1A-1B的主要层104或次要层106)中的一系列WC。主要阶段(S1)是数据输入阶段或其中刺激(例如,输入IWC 402)被输入到系统的地方。在一些实施方式中,通过引入除了输入数据电压I_{WC}之外的电压,使用权重电压(WV)电路来控制到WC的输入权重。在各种实施方式中,不同WC(例如,WV 404、WV 406和WV 408)的权重电压可以被设置为相同或不同的电压值。在一些实施方式中,电路(本文中可以被称为加权输入模块)用于将输入刺激与权重电压组合,并向各个WC供应净加权输入。例如,在图4中,电路440将来自WC 404的权重电压与输入刺激I_{WC} 402组合以向比较器WC_1 450供应加权输入,并且电路442将来自WC 406的权重电压与输入刺激I_{WC} 402组合以向比较器WC_2 452供应加权输入,并且电路444将来自WC 408的权重电压与输入刺激I_{WC} 402组合以向比较器WC_N 454供应加权输入。在一些实施方式中,子分类器包括耦合到比较器模块的加权输入模块。例如,在图4中,比较器WC_1 450被耦合到加权输入模块440,比较器WC_2 452被耦合到加权输入模块442,比较器WC_N 454被耦合到加权输入模块444。

[0056] 对于一些拓扑结构,根据一些实施方式,锁存器或临时存储器单元(本文中有时称为存储器模块,例如,锁存器410、412和414)的次要阶段(S2)电路存储WC(O_{WC})的输出。存在使用不同的序列来存储数据的各种锁存技术。例如,在简化的序列中,锁存器在指定的时间段内一次存储对一个刺激的响应。在更复杂的序列中,输入数据被划分为数据块,并且锁存器一次存储对应于一个数据块的响应数据。

[0057] 在一些实施方式中,根据一些实施方式,基于神经网络的拓扑或互联性的方法,将各个WC的输出(O_{WC})数据输入到其他WC。在一些实施方式中,如图4中所示,放置在次要阶段(S2)存储器锁存器之后的神经网络中的另一电路的集合(例如,电路416、418和420)将电压馈送回相邻神经元(WC),以便调整阶段3(S3)中WC的输入阶段的WV。在一些实施方式中,来自神经网络中的一个或多个神经元的附加连接(例如,连接428、430和432)被施加于由拓扑总线O_{NN}(例如,拓扑总线422、424或426)指示的WV电路。在一些实施方式中,在没有锁存器的次要阶段(S2)电路的情况下,可以设置用于神经元的激活的时钟周期或脉冲宽度,使得电流反馈不会使神经元网络饱和,并且神经元仅基于神经元的输出而启动。

[0058] 尽管拓扑总线(例如,总线422、424和426)在图4中由同一名称O_{NN}标识,但在各种实施方式中,拓扑总线被不同地配置;例如,神经网络的不同部分或区域可以基于局部拓扑被不同地组织或互连。类似于人脑被划分为高度专业化的区域的方式,在一些实施方式中,硅

的单晶片(具有数百万甚至数十亿个WC)可以进一步被细分为神经网络的区域,具有相互连接形成人工脑的专门用于特定功能(诸如语音、视觉、听觉等)个人拓扑。

[0059] 由于电路的类型和硬件连接,所有WC都可以以并行计算的方式同时处理,其与完全基于软件的网络相比,产生显著的性能增益,同时还提供广泛的应用。例如,具有蚀刻在半导体晶片上的一百万个WC,一百万个或更多WC的整个集合可以在单个时钟周期内被评估。假设时钟频率为2GHz,例如,可以在一秒钟内评估一百万个或更多个神经元的神经网络(WCs)的20亿次或更多次迭代。

[0060] 为了进一步说明如何使用上面参考图2-4所述的WC网络来建立神经网络,根据一些实施方式,考虑自组织映射(SOM)——无监督学习网络的示例构造。SOM中的自组织处理初始化、竞争、协作和适应组成。最初,利用小的随机值初始化每个神经元的权重向量。在WC硬件神经元的情况下,权重电压(WVs)被初始化为随机值。神经元计算它们的用于每个输入模式(例如,图像中的所有像素)的判别函数的各个值。通常用在传统的虚拟神经元中的判别函数是每个神经元的输入向量和互联权重向量之间的平方欧氏距离。对于基于WC的硬件神经元,判别函数可以是输入电压和权重电压之间的平方欧氏距离。具有最小的判别函数值的特定神经元被认为是决定神经元。对于基于WC的硬件神经网络,取决于电压阈值的初始配置和相关的权重,一个或多个WC神经元可以是响应于输入模式的决定神经元。为了简单起见,例如考虑单个决定神经元。决定神经元基于拓扑结构确定协作的激活神经元的邻域的空间位置(例如,以调整邻域中的权重)。当一个神经元被激活时,它最邻近的邻居比那些位于更远处的那些邻居更容易受到影响。在基于WC的硬件神经网络中,由于来自决定神经元的电压输出与其相邻居相连接,所以输出影响邻居的权重电压。受影响的神经元通过相关的连接权重的调整来更新它们的与输入模式有关的判别函数的个别值。在基于WC的神经网络中,权重是连续自适应的。决定神经元对随后的类似输入模式的应用的响应因此增强。

[0061] 作为可视化基于WC的神经网络的自组织过程的方法,考虑如何将在连续二维输入空间中输入的数据的集合映射到基于WC的神经元的集合上。根据拓扑来组织或连接基于WC的神经元(例如,每个神经元与每个其他的神经元相连)。基于WC的神经元可以从随机分配(例如,电压值)开始,并且权重被初始化为随机初始值或根据等级。每个神经元读取被转换(例如通过预处理器)成相应的电压值的第一输入。神经元中的一个——“决定神经元”——将以高值输出进行响应。在不同的配置中,多于一个神经元可以对输入进行响应。决定神经元被称为向数据输入移动,因为决定神经元的权重的初始值响应于输入电压而被调整,以响应于输入电压具有决定神经元及其邻居。相邻的神经元也向数据输入移动,但移动较小量。因为所有的神经元在每一步馈送相同的输入来选择一个或多个决定神经元和/或相关的邻居,所以该过程是并行的(即,基于WC的神经元一致地操作)。权重(所有基于WC的神经元的电压值)在该步骤结束时被调整。接下来,选择第二个数据输入以进行训练。与第一“决定神经元”不同的神经元在第二轮是决定性的。并且与新决定神经元紧邻的神经元通过向第二个输入数据移动较小的量而进行响应。在该步骤结束时,再次调整权重。该过程持续到所有神经元的权重达到稳定状态(例如,神经网络中神经元的权重电压不再有大的变化),并且至少直到所有输入数据被处理。例如,使用给定的数据集合多次重复该过程。最后,基于WC的神经元的整个输出网格表示输入空间。

[0062] 图5示出了具有来自CG模块(未示出)的添加的CG信号502的一系列WC(如图4所示)。根据一些实施方式,CG模块控制神经网络的神经可塑性和行为。利用神经可塑性,神经网络执行自学习。传统的网络经由重复暴露于数据集并在一段时间内收敛来训练,并且调整连接的权重来匹配数据集以产生有意义的输出。训练时间段取决于学习速率。例如,对于自组织映射(SOM),训练时间段还取决于被称作训练的影响的学习半径或神经元数量。学习半径指示距最佳匹配单元(BMU)——为特定输入激活的神经元(本文中有时称为"决定神经源")——的距离。这些参数逐渐减小,直到神经网络被完全训练来以期望的方式响应于刺激。然而,在初始训练期间没有考虑到的任何新的数据集落在训练神经网络的范围之外。这限制了神经网络实现神经可塑性的能力,并且神经网络必须被再训练以处理新的数据集或被重新设计以适应新的数据。

[0063] 为了解决传统神经网络的这些局限性,在一些实施方式中,如图5所示,CG信号(例如,信号502)增强或抑制对输入刺激的响应,并修改学习速率和/或神经元数量。在一些实施方式中,CG信号被持续和/或间歇地施加于神经网络,以影响神经网络的行为。这种行为变化包括神经网络响应于刺激是如何专注、放松和/或反应。在一些实施方式中,CG信号将神经网络的焦点限制于特定的刺激和/或神经网络的整体学习能力。即使在神经网络处理刺激时,CG信号因此实现了自适应学习。在一些实施方式中,同时使用多于一个CG变量,覆盖复杂的行为模式。在一些实施方式中,CG信号可以影响一组WC神经元的局部行为/可塑性和/或全局行为/可塑性。在一些实施方式中,CG信号可以影响比较模块的灵敏度。在一些实施方式中,CG信号可以先于加权模块影响输入信号。在图5中,例如,CG信号502影响到神经元或比较器WC_1、WC_2和WC_N的输入。在一些实施方式中,不同的CG信号被施加于WC神经元的不同区域或邻域以影响神经网络行为。

[0064] 根据一些实施方式,WV电路(例如,WV 504、WV 506、WV 508)接收CG信号(例如,信号502),并基于CG信号调整对各个WC的加权输入。在各种实施方式中,WV电路使用电压控制电阻器(VCR)和/或可变电阻器(例如,电位计或数字电位计、场效应晶体管、电阻器梯、电阻器桥、电阻器网络、结晶体管或其他电流或电压控制电路)来构建,其取决于CG信号,通过WC控制与输入刺激相比较的加权输出。

[0065] 图6示出了根据一些实施方式,具有图5的CG信号的一系列WC和附加存储器604(本文中也称为隐式存储器(IMEM))。IMEM 604可以允许存储器指针到子区域的快速重定向,和/或为包含响应刺激所必需的数据的存储器区域提供存储器地址。用IMEM 604,(图1A-1B的)评估系统116避免读取大的存储器空间来搜索特定于给定输入刺激的数据,以便评估特定数据并提供响应114。

[0066] 根据一些实施方式,隐式存储器可以增强神经网络以具有对刺激的直观反应,例如,通过即使输入刺激只类似于(而且不完全匹配)以前的经验,也触发响应。与IMEM相比,根据一些实施方式,显式存储器块(例如,EMEM块606、608和610)可以被配置成存储响应于输入刺激而要被检索的精确数据(例如,针对过去输入的过去响应)。例如,PNP AI处理器102可以将当前输入与先前输入相匹配(例如,相当于已经访问过房间或看过视频或图像的人),可以从EMEM检索先前生成的虚拟图像,并将其与当前输入相比较以生成匹配响应。更详细的数据可以经由EMEM访问,而IMEM存储并表示从数据中提取的信息和概念的一般模式。

[0067] 在一些实施方式中,存储器604可以被可视化为一个或多个存储器块的集合,其中每个存储器块表示要被检索的数据。根据一些实施方式,内存块可以被引用作为IMEM块,也可以作为EMEM块。在一些实施方式中,IMEM使用一个或多个WC输出(例如,来自神经网络的给定神经元块的输出)和/或CG状态的组合是可寻址的。例如,在图6中,IMEM 604经由控制信号602(分别由与相应的EMEM块606、608和610连接的信号618、620和622示出)和/或WC输出612、614和616可直接寻址。在一些这样的实施方式中,CG信号影响存储器块的大小,或为响应刺激而访问的存储器块的数量。与IMEM块相反,EMEM块经由WC输出寻址。例如,根据一些实施方式,使用连接线(例如,线612、614和616)来寻址图6中的EMEM块。

[0068] 如图6所指示的,根据一些实施方式,来自不同存储块(例如,块624、626和628)的数据被评估系统116(在图1A-1B示出)使用以响应输入刺激(例如,输出以响应过程624、626和628)。

[0069] 与缓存内存器架构通过先前存储和/或经常使用的功能或数据以用于更快的访问的方式类似,IMEM架构基于对刺激的熟悉程度来改进内存访问。例如,反复观察的刺激可以向系统提供反馈,使得一个或多个控制组信号可以用于直接访问存储器中的一个或多个对象,而不必依赖于模式匹配神经网络的输出来指定存储器位置。但是,与缓存不同,IMEM改进存储器访问(例如,经由使用CG和WC输出的直接访问),而不需要额外的存储或执行重复的搜索迭代来解析和找到正确的内存位置。

[0070] 尽管图6示出了EMEM块,如各个EMEM块连接到仅单个子分类器(例如,将第一个子分类器连接到EMEM块1的线612),子分类器可以访问多于一个EMEM块。连接线(例如,线612、614和616)旨在经由比较器的输出来显示EMEM块的可寻址性。

[0071] 如块630所指示的,在一些实施方式中,主分类器包括多个子分类器、IMEM块和/或CG信号。在一些实施方式中,主分类器630是耦合到多个子分类器、IMEM块和/或CG信号的独立模块(图6中未示出)。在一些这样的实施方式中,主分类器630在一个或多个时钟周期期间(例如,经由存储器604)从多个子分类器中的每个接收各个输出信号中的每个,并且基于产生具有大于预定输出阈值的值的各个输出信号的多个子分类器的子集来确定分类器响应。

[0072] 图7是示出根据一些实施方式的具有并行神经处理器(例如,PNP AI处理器102)的代表性系统700的框图。在一些实施方式中,系统700(例如,具有系统架构100的任何设备,图1A)包括一个或多个处理单元(例如,CPU、ASIC、FPGA、微处理器等)702、一个或多个通信接口714、存储器718、外部传感器704、音频/视频输入706、以及用于互连这些组件(有时称为芯片组)的一个或多个通信总线720。

[0073] 在一些实施方式中,系统700包括接口708。在一些实施方式中,用户接口708包括能够呈现媒体内容的一个或多个输出设备710,其包括一个或多个扬声器和/或一个或多个视觉显示器。在一些实施方式中,用户接口708还包括一个或多个输入设备712,其包括有助于用户输入的用户界面组件,诸如键盘、鼠标、语音命令输入单元或麦克风、触摸屏显示器、触敏输入板、手势捕捉相机或其他输入按钮或控件。此外,一些系统使用麦克风和语音识别或相机和手势识别来补充或替代换键盘。

[0074] 在一些实施方式中,系统700包括一个或多个图像/视频捕获或音频/视频输入设备706(例如,照相机、摄像机、扫描仪、照片传感器单元)。可选地,系统700包括用于确定系

统设备700的位置的位置检测设备(未示出)。

[0075] 在一些实施方式中,系统700包括一个或多个内置传感器718。在一些实施方式中,内置传感器718包括例如一个或多个热辐射传感器、环境温度传感器、湿度传感器、IR传感器、占用传感器(例如,使用RFID传感器)、环境光传感器、运动检测器、加速计和/或陀螺仪。

[0076] 在一些实施方式中,系统700包括一个或多个外部传感器704。在一些实施方式中,外部传感器704包括例如一个或多个热辐射传感器、环境温度传感器、湿度传感器、IR传感器、占用传感器(例如,使用RFID传感器)、环境光传感器、运动检测器、加速计和/或陀螺仪。

[0077] 根据一些实施方式,系统700包括用于执行/分流上面参考图1-6描述的AI任务的一个或多个并行神经处理器716(例如,图1A或1B中的PNP AI处理器102)。

[0078] 通信接口720例如包括能够使用各种定制或标准无线协议(例如,IEEE 802.15.4、Wi-Fi、ZigBee、6LoWPAN、线程、Z-波、智能蓝牙、ISA100.11a、无线HART、MiWi等)中的任何一个和/或各种定制或标准有线协议(例如,以太网、家庭插头等)或包括截至本文件申请日尚未开发的通信协议的任何其他适当的通信协议。

[0079] 存储器721包括高速随机存取存储器,诸如DRAM、SRAM、DDR RAM或其他随机存取固态存储器设备;以及,可选地包括非易失性存储器,诸如一个或多个磁盘存储设备、一个或多个光盘存储设备、一个或多个闪存设备,或一个或多个其他非易失性固态存储设备。存储器721或者可替选地存储器721内的非易失性存储包括非暂时性计算机可读存储介质。在一些实施方式中,存储器721或存储器721的非暂时性计算机可读存储介质存储以下程序、模块和数据结构或其子集或超集:

[0080] • 操作逻辑722,该操作逻辑722包括处理各种基本系统服务和执行硬件相关任务的程序;

[0081] • 设备通信模块724,该设备通信模块724用于连接到经由一个或多个通信接口720(有线的或无线的)连接到一个或多个网络的其他网络设备(例如,网络接口,诸如提供互联网连接的路由器、网络存储设备、网络路由设备、服务器系统等)并且与其他网络设备通信;

[0082] • 输入处理模块726,该输入处理模块726用于检测来自一个或多个输入设备712的一个或多个用户输入或交互,并解释检测到的输入或交互;

[0083] • 用户接口模块728,该用户接口模块728用于提供和显示其中可以配置和/或查看一个或多个设备(未示出)的设置、捕获的数据和/或其他数据的用户界面;

[0084] • 一个或多个应用模块730,该一个或多个应用模块730由系统700执行,用于控制设备,并用于复查由设备捕获的数据(例如,设备状态和设置、捕获的数据或关于系统700和/或其他客户端/电子设备的其他信息);

[0085] • PNP预处理模块732,该PNP预处理模块732为PNP AI处理器716提供预处理数据的功能,包括但不限于:

[0086] o数据接收模块7320,该数据接收模块7320用于从一个或多个输入设备712、外部传感器704、内置传感器718和/或音频/视

[0087] 频输入706接收要由PNP AI处理器716处理的数据;

[0088] o数据预处理模块7322,该数据预处理模块7322用于处理由数据接收模块7320捕获或接收的数据,并用于准备(例如,用于从原始数据输入创建向量的集合,将向量组织成

表,和/或将原始数据转换成电压值),以及用于将经处理的数据(例如,通过施

[0089] 加电流加载数据)发送到PNP AI处理器716;

[0090] • PNP训练模块734,该PNP训练模块734与PNP预处理模块732和/或PNP反馈和响应模块734(如下所述)协调,以训练一个或多个PNP AI处理器716(例如,设置电压值/阈值、初始化神经网络以及监控学习速率和进展);以及

[0091] • PNP反馈和响应模块734,包括但不限于:

[0092] o数据接收模块7360,该数据接收模块7360用于从PNP AI处理器716接收数据(例如,用于从窗口比较器电路的输出接收

[0093] 电压值);

[0094] o数据后处理模块7362,该数据后处理模块7362用于后处理从PNP AI处理器716接收的数据(例如,用于将电压值或神经网

[0095] 络输出转换为由系统进一步处理有用的另一格式);

[0096] o反馈模块7364,该反馈模块7364用于基于PNP AI处理器716的输出(例如,用于重新调整控制值)或基于来自系统中的其他设备的输入(包括改变的环境)来生成对PNP AI处理器716的

[0097] 反馈;以及

[0098] o响应模块7366,该响应模块7366用于基于PNP AI处理器的输出生成系统响应(例如,移动机械臂、改变照相机的位置或用信令发送警报)。

[0099] 上述识别的元素的每个可以存储在一个或多个先前提到的存储器设备中,并且对应于用于执行上述功能的指令的集合。上述识别的模块或程序(即指令的集合)不需要作为单独的软件程序、过程或模块来实现,并且因此在各种实施方式中这些模块的各种子集可以组合或以其他方式重新排列。在一些实施方式中,存储器606可选地存储以上标识的模块和数据结构的子集。此外,存储器606可选地存储以上未描述的附加模块和数据结构。

[0100] 图8A-8D示出了根据一些实施方式,使用包括加权输入模块和耦合到加权输入模块的比较模块的子分类器处理输入信号(802)的方法800的流程图表示。以上参考图4描述了子分类器。在一些实施方式中,比较模块包括(804)至少一个运算放大器模块(例如,以上参考图2A和2B描述的模拟WC)。第一子分类器在加权输入模块处对输入信号施加(806)权重以生成加权输入信号。例如,在图4中,加权输入模块440对输入信号402施加权重404,以生成到比较器WC_1 450的加权输入信号。在一些实施方式中,加权输入模块从第二子分类器接收(808)第二输出信号。例如,在图4中,加权输入模块442从其他子分类器(例如,从来自存储器锁存器410输出的子分类器)接收输出信号418。

[0101] 方法800进一步包括在比较模块处,在比较模块输入线处接收(810)加权输入信号。例如,在图4中,比较器WC_1 450在所示的将模块440与比较器450连接的线上从加权输入模块440接收加权输入信号。在一些实施方式中,比较模块包括至少一个运算放大器模块(以上参考804中描述),并且接收加权输入信号包括在至少一个运算放大器模块处接收(812)加权输入信号。

[0102] 如图8B所示,根据一些实施方式,方法800进一步包括在比较模块处,在比较模块输出线处生成(814)第一输出信号。例如,在图4中,比较器WC_1 450在将比较器WC_1 450与存储器锁存器410连接的线上生成输出信号0^1_{WC}。生成第一输出信号(814)包括在比较模块处

确定(816)是否加权输入信号具有在下窗口范围值和上窗口范围值之间的值。根据一些实施方式,以上参照图2A、2B、3和4描述了比较操作。例如,在图2A中,Op-Amp 202和Op-Amp 204确定输入电压I_{WC} 206是否在较低电压阈值R_L 208和较高电压阈值R_H 210之间。生成第一输出信号(814)进一步包括响应于加权输入信号具有在下窗口范围值和上窗口范围值之间的值的确定,在比较模块处,在比较模块输出线处,将第一输出信号设置(818)具有第一值。例如,在图2A中,如果输入电压I_{WC} 206在较低电压阈值R_L 208和较高电压阈值R_H 210之间,则Op-Amp 204设置输出O_{WC}。作为另一示例,在图4中,如果输入电压I_{WC} 402在较低电压阈值$R_L{}^1$和较高电压阈值$R_H{}^1$之间,比较器WC_1(450)在连接比较器和存储器锁存器410的线处设置输出$O^1{}_{WC}$(为高电压值)。生成第一输出信号(814)进一步包括响应于加权输入信号具有不在下窗口范围值和上窗口范围值之间的值的确定,在比较模块处,在比较模块输出线处将第一输出信号设置(820)成具有与第一值不同的第二值。例如,在图2A中,如果输入电压I_{WC} 206不在较低电压阈值R_L 208和较高电压阈值R_H 210之间,则Op-Amp 202设置输出/O_{WC}。作为另一示例,在图4中,如果输入电压I_{WC} 402在较低电压阈值$R_L{}^1$和较高电压阈值$R_H{}^1$之间,则比较器WC_1(450)在连接比较器和存储器锁存器410的线处设置输出$O^1{}_{WC}$(为低电压值)。在一些实施方式中,比较模块包括至少一个运算放大器(例如,如图2A所述),并且生成第一输出信号(814)进一步包括在至少一个运算放大器模块设置(822)第一输出信号。

[0103] 在一些实施方式中,如图8C所示,方法800进一步包括将第一输出信号从第一子分类器发送(824)到第二子分类器。例如,在图4中,来自比较器WC_1 450、被锁存在存储器锁存器410中的输出信号被发送(如虚线、输入418所示)到包括比较器WC_2 452和加权输入模块442的第二子分类器。

[0104] 在一些实施方式中,方法800进一步包括在加权输入模块处接收(826)控制组信号。在一些这样的实施方式中,加权输入模块包括可变电阻器模块,并且该方法包括在可变电阻器模块处接收(828)控制组信号并且基于控制组信号来调整加权输入信号。在一些实施方式中,方法800进一步包括在加权输入模块处对输入信号施加(830)权重以基于控制组信号生成加权输入信号。以上参照图5说明了接收和处理控制组信号。例如,在图5中,在被相应的加权输入模块消耗之前,控制组信号502被施加并且调整电路504、506、508中的权重值。

[0105] 在一些实施方式中,如图8D所示,第一子分类器包括(832)耦合到比较模块的存储器模块。在一些实施方式中,存储器锁存器、比较器和加权输入模块包括子分类器。在一些这样的实施方式中,第一子分类器在存储器模块从比较模块接收第一输出信号并存储(834)第一输出信号,以及向第二子分类器提供第一输出信号。例如,在图4中,根据一些实施方式,存储器锁存器410耦合到比较器WC_1(450),存储器锁存器412耦合到比较器WC_2(452),并且存储器锁存器414耦合到比较器WC_N(454)。

[0106] 图9A-9E示出了使用分类器系统在一个或多个时钟周期期间处理一个或多个输入信号的方法900的流程图表示。根据一些实施方式,分类器系统包括(902)多个子分类器和耦合到多个子分类器的主分类器,多个子分类器均包括加权输入模块和比较模块。以上参考图6讨论了耦合到多个子分类器的示例主分类器。在方法900的一些实施方式中,如图9B中920所示,子分类器中的每个具有在下窗口范围值和上窗口范围值之间、不与任何其他子分类器的任何其他各个窗口范围重叠的各个窗口范围。例如,在图3中,比较器WC_1的参考电

压($R_L^1$302和R_H^1 304)、比较器WC_2的参考电压(R_L^2 306和R_H^2 308)、比较器WC_3的参考电压(R_L^3 310和R_H^3 312)和比较器WC_N的参考电压(R_L^N 314和R_H^N 316)每个被设置成使得对应的WVR都是唯一的。在一些实施方式中,WVR是不重叠的。在方法900的一些实施方式中,如图9C中922所示,分类器系统包括耦合到多个子分类器的复用器模块。在一些这样的实现中,方法900包括在复用器模块处,在单个时钟周期期间向多个子分类器提供(924)输入信号中的一个。以上参考图1A描述了复用器和/或预处理模块的示例。

[0107] 在一些实施方式中,方法900包括(906)在每个子分类器处,在加权输入模块处施加(908),针对用于各个时钟周期的各个输入信号施加权重以生成加权输入信号。在一些这样的实施方式中,方法(900)进一步包括在每个子分类器处,在比较模块处,在比较模块输入线处接收(910)加权输入信号。在一些这样的实施方式中,方法(900)进一步包括在比较模块处,在比较模块输出线处生成(912)输出信号。根据一些实施方式,以上参考图4描述了示例加权输入模块。

[0108] 在一些实施方式中,如图9D所示,在比较模块生成(912)输出信号包括:在比较模块处,确定(926)是否加权输入信号具有在下窗口范围值和上窗口范围值之间的值;响应于确定加权输入信号具有在下窗口范围值和上窗口范围值之间的值,在比较模块处,在比较模块输出线处将输出信号设置(928)成具有大于预定输出阈值的值;以及响应于确定加权输入信号具有不在下窗口范围值和上窗口范围值之间的值,在比较模块处,在比较模块输出线处将输出信号设置(930)成具有小于预定输出阈值的值。根据一些实施方式,以上参考图8B描述了比较模块的操作,并且应用于图9D所示的操作。

[0109] 返回参考图9A,在一些实施方式中,方法900包括(914)在主分类器处,在一个或多个时钟周期内从多个子分类器中的每个接收(916)输出信号中的每个。在一些这样的实施方式中,方法900进一步包括:在主分类器处,基于产生具有大于预定输出阈值的值的各个输出信号的多个子分类器的子集来确定(918)分类器响应。

[0110] 在方法900的一些实施方式中,分类器包括(932)耦合到主分类器和/或多个子分类器的存储器块。在一些这样的实施方式中,该方法进一步包括在多个子分类器(934)的每个的加权输入模块处:接收(936)控制组信号,并基于控制组信号对输入信号施加(938)权重以生成加权输入信号。该方法进一步包括在存储器块存储(940)分类器的一个或多个响应,存储器块包括使用多个子分类器的一个或多个子分类器的输出信号和控制组信号可寻址的一个或多个存储器子块。该方法进一步包括在主分类器处基于一个或多个存储器子块来确定(942)分类器响应。根据一些实施方式,以上参考图6讨论了主分类器的细节。此外,如图6相关讨论中所指示的,根据一些实施方式,使用EMEM块的输出(例如,输出624、626和628)来确定分类器响应。

[0111] 应当理解,图8A-8D和9A-9E中的操作已经描述的操作的特定次序仅是示例,并不旨在指示所描述的次序是可以执行的操作的唯一次序。一个本领域的普通技术人员将认识到对本文中描述的操作重新排序的各种方式。此外,应注意,关于方法800所描述的其他过程的细节也以与以上关于图9A-9E所描述的方法900类似的方式适用。

[0112] 尽管各种附图中的一些以特定次序示出了多个逻辑阶段,但是可以重新排序不依赖于顺序的阶段,并且可以组合或分解其他阶组曲。虽然特别提及了一些重新排序或其他分组,但是对于本领域普通技术人员来说,其他的将是显而易见的,因此本文提出的排序和

分组不是可选方案的详尽列表。此外,应当认识到,这些阶段可以在硬件、固件、软件或其任何组合中实现。

[0113] 为了说明的目的,已经参考具体实施方式描述了上述描述。然而,以上说明性讨论不旨在详尽无遗或将权利要求的范围限于所公开的确切形式。鉴于以上教导,许多修改和变化是可能的。选择这些实施方式以便最好地解释权利要求书的基本原理及其实际应用,从而使本领域技术人员最好地使用具有适合于预期的特定用途的各种修改的实现。

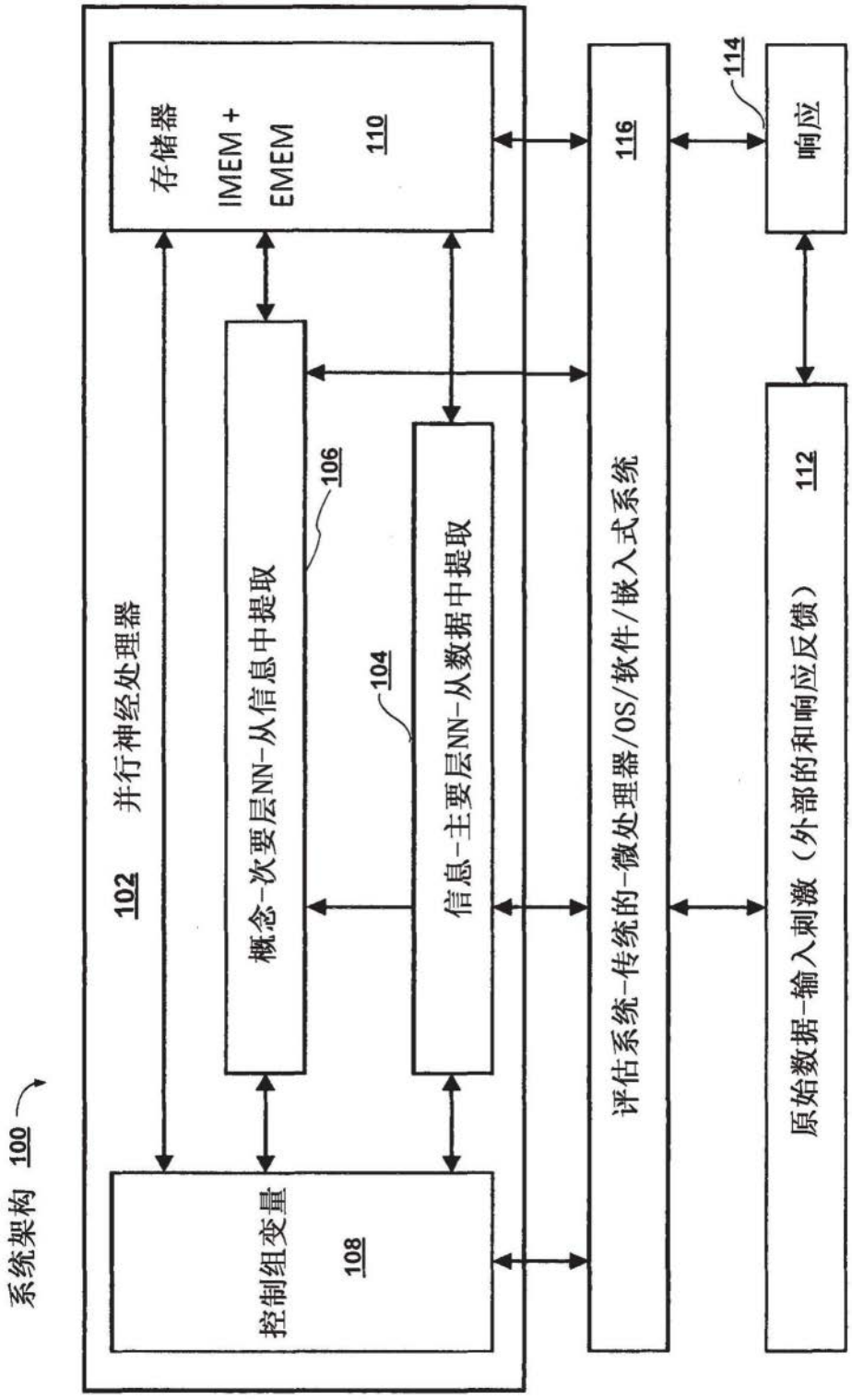

系统架构 100
102 并行神经处理器
存储器
IMEM +
EMEM
110
概念-次要层NN-从信息中提取
106
信息-主要层NN-从数据中提取
104
控制组变量
108
评估系统-传统的-微处理器/OS/软件/嵌入式系统
116
114
响应
原始数据-输入刺激（外部的和响应反馈）
112

图1A

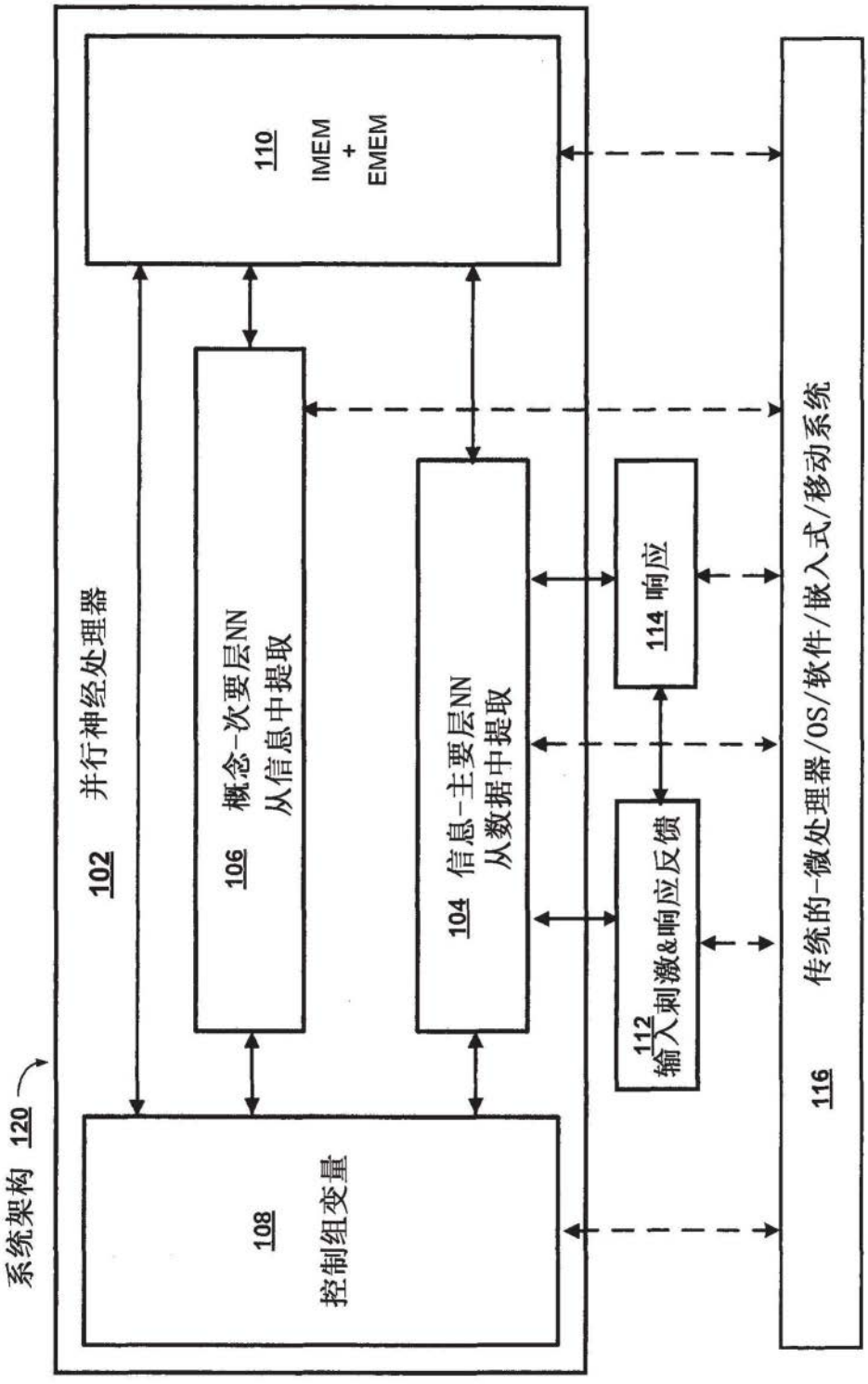
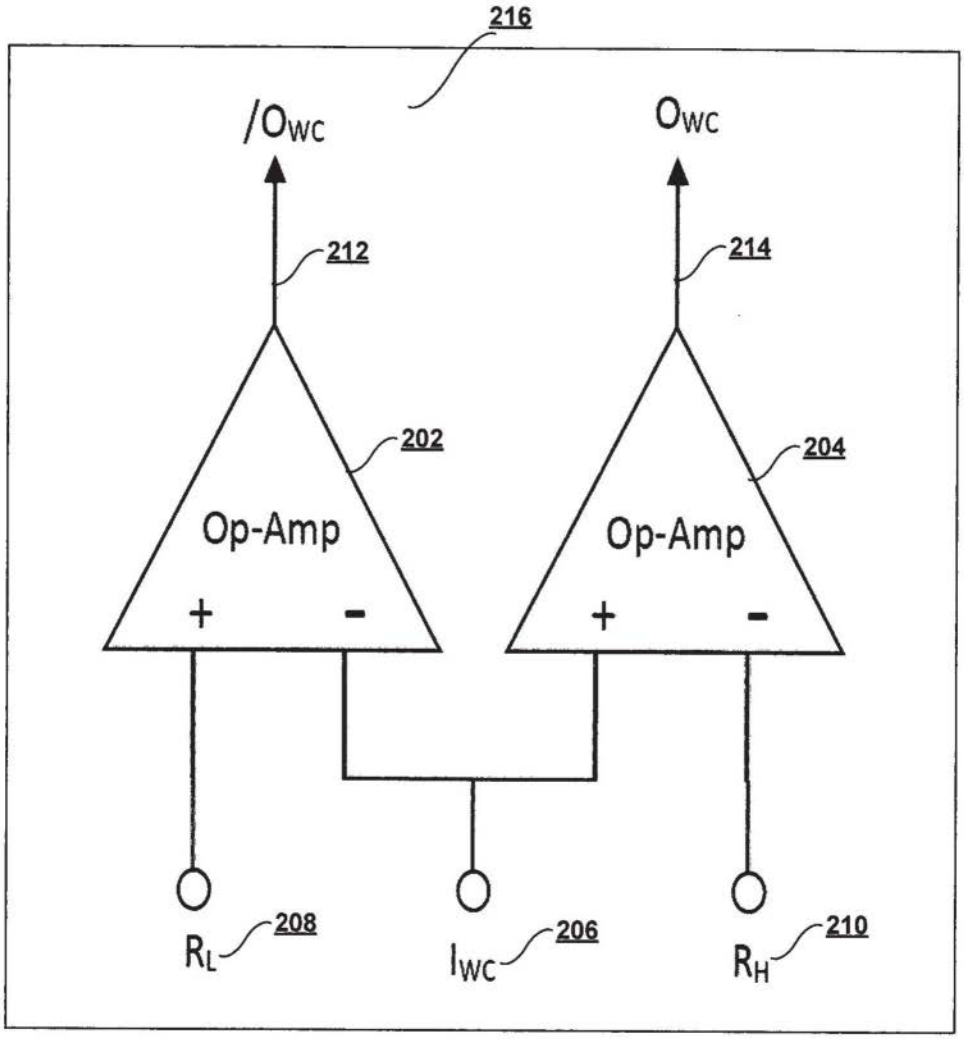

系统架构 120
102 并行神经处理器
110 IMEM + EMEM
106 概念-次要层NN 从信息中提取
104 信息-主要层NN 从数据中提取
108 控制组变量
112 输入刺激&响应反馈
114 响应
116 传统的-微处理器/OS/软件/嵌入式/移动系统

图1B



O_{WC} = 高 = 1 $(R_L < I_{WC} < R_H)$

$/O_{WC}$ = 低 = 0 $(I_{WC} < R_L)$ or $(I_{WC} > R_H)$

图2A

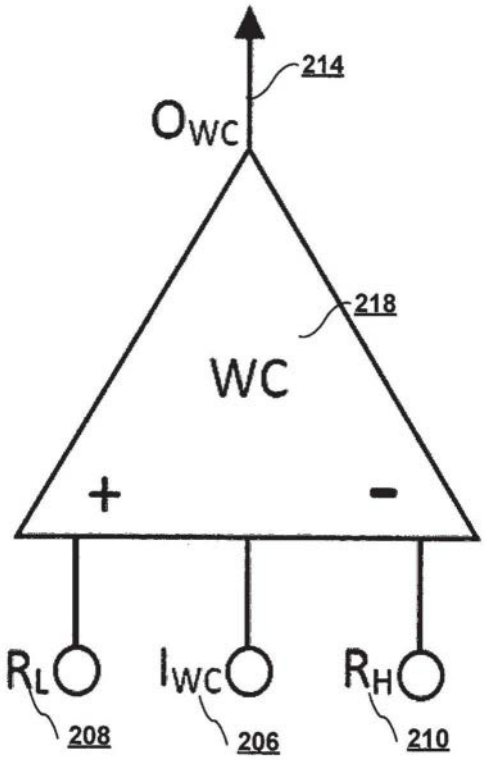


图2B

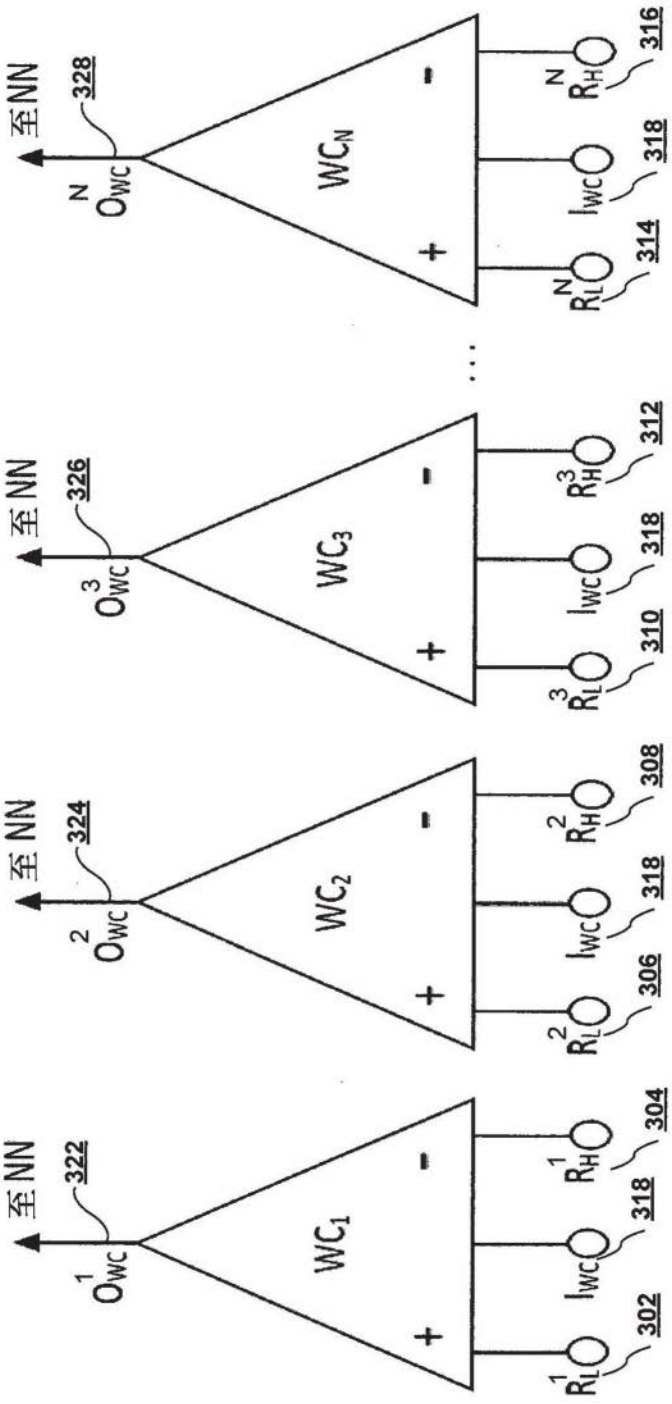

至NN
322
O^1_{WC}
WC_1
R^1_L
I_{WC}
R^1_H
302
318
304
至NN
324
O^2_{WC}
WC_2
R^2_L
I_{WC}
R^2_H
306
318
308
至NN
326
O^3_{WC}
WC_3
R^3_L
I_{WC}
R^3_H
310
318
312
至NN
328
O^N_{WC}
WC_N
R^N_L
I_{WC}
R^N_H
314
318
316

图3

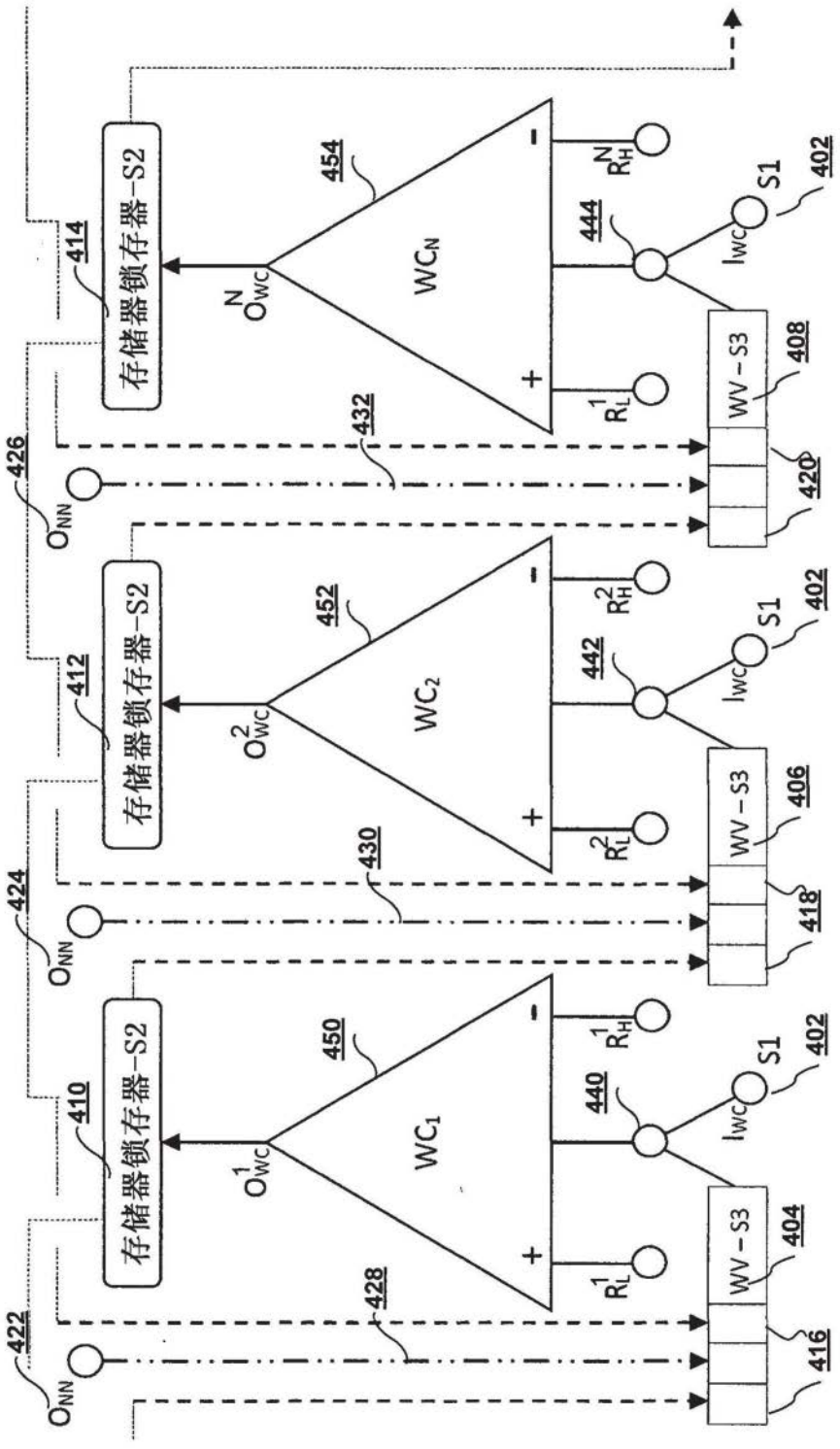

存储器锁存器-S2
存储器锁存器-S2
存储器锁存器-S2
WV-S3
WV-S3
WV-S3
S1
S1
S1
402
402
402
404
406
408
410
412
414
416
418
420
422
424
426
428
430
432
440
442
444
450
452
454
WC1
WC2
WCN

图4

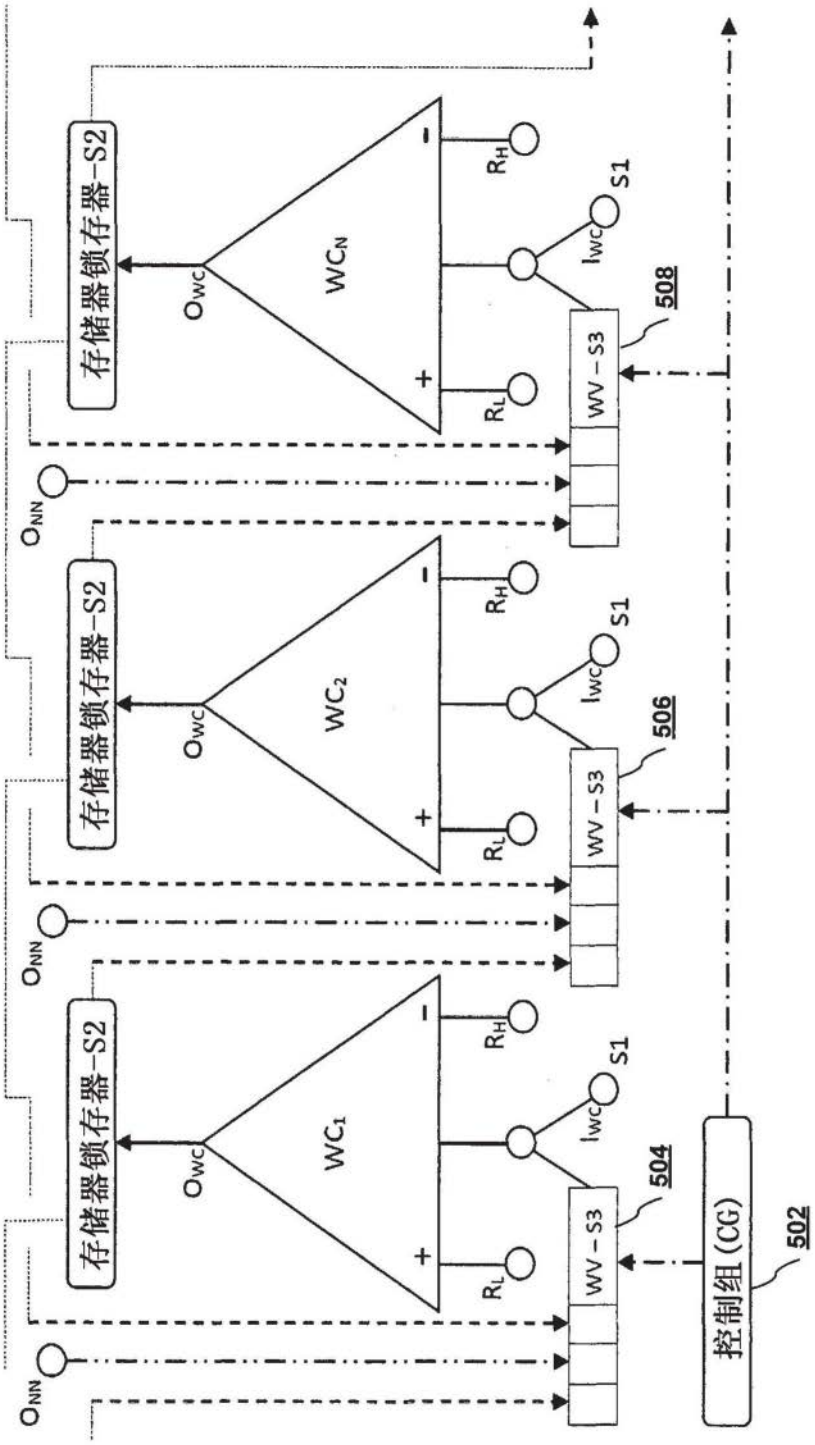

存储器锁存器-S2
存储器锁存器-S2
存储器锁存器-S2
WC_N
WC_2
WC_1
O_{WC}
R_H
R_L
I_{WC}
S1
WV-S3
504
506
508
O_{NN}
控制组(CG)
502

图5

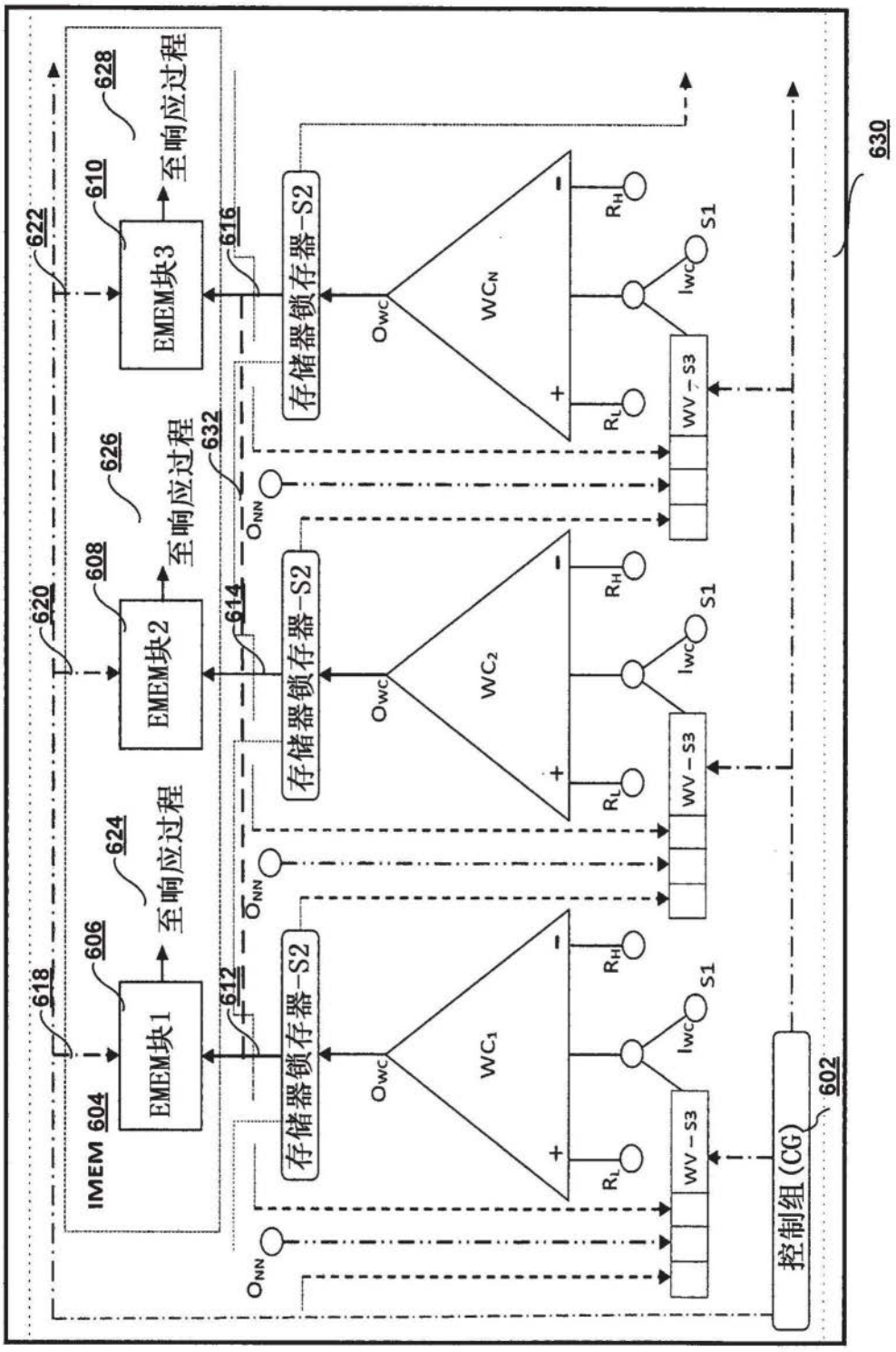

IMEM 604
EMEM块1
EMEM块2
EMEM块3
至响应过程
至响应过程
至响应过程
存储器锁存器-S2
存储器锁存器-S2
存储器锁存器-S2
WC1
WC2
WCN
WV-S3
S1
控制组(CG)
602
606
608
610
612
614
616
618
620
622
624
626
628
630
632

图6

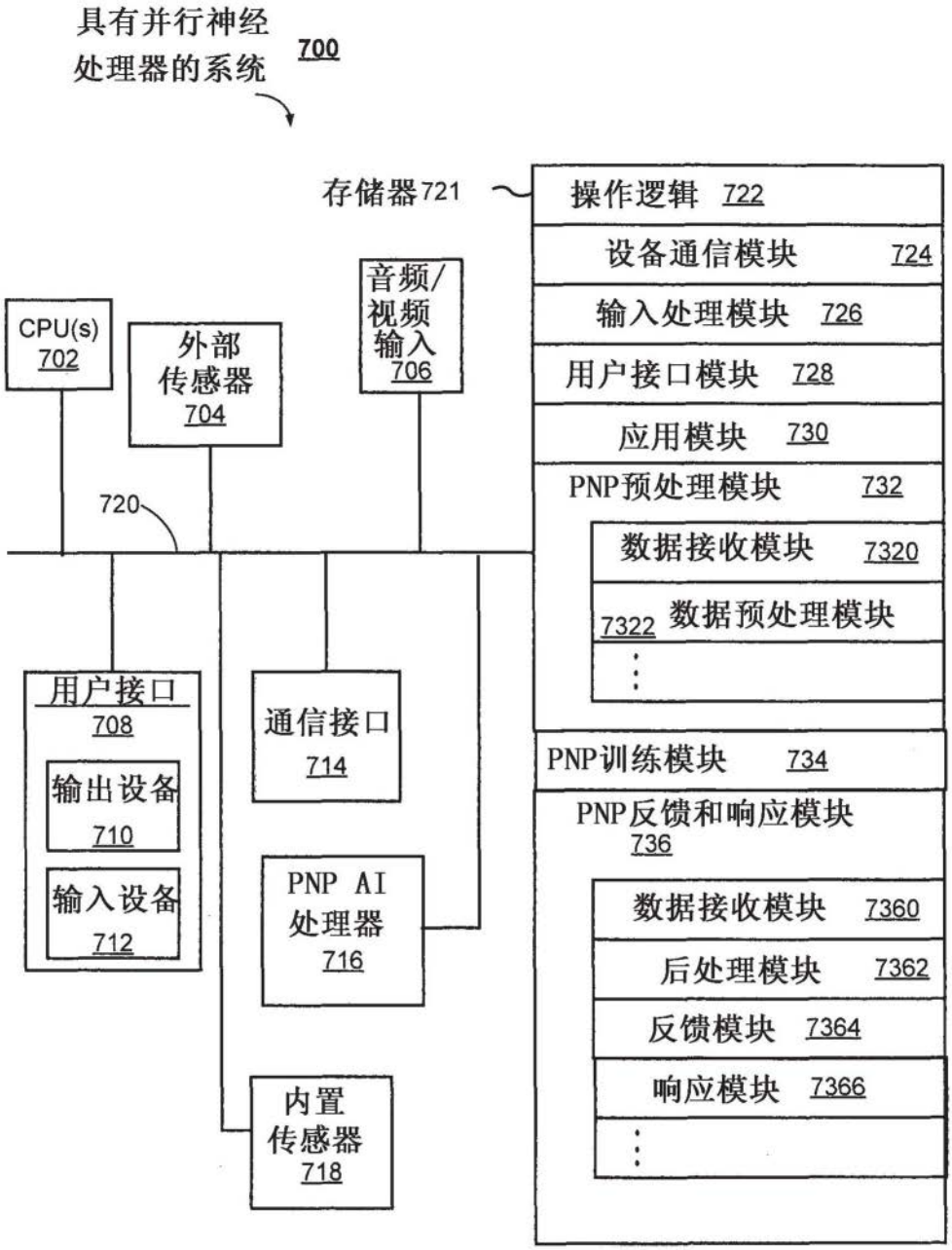

具有并行神经处理器的系统 700
存储器721
操作逻辑 722
设备通信模块 724
输入处理模块 726
用户接口模块 728
应用模块 730
PNP预处理模块 732
数据接收模块 7320
7322 数据预处理模块
PNP训练模块 734
PNP反馈和响应模块 736
数据接收模块 7360
后处理模块 7362
反馈模块 7364
响应模块 7366
CPU(s) 702
外部传感器 704
音频/视频输入 706
720
用户接口 708
输出设备 710
输入设备 712
通信接口 714
PNP AI处理器 716
内置传感器 718

图7

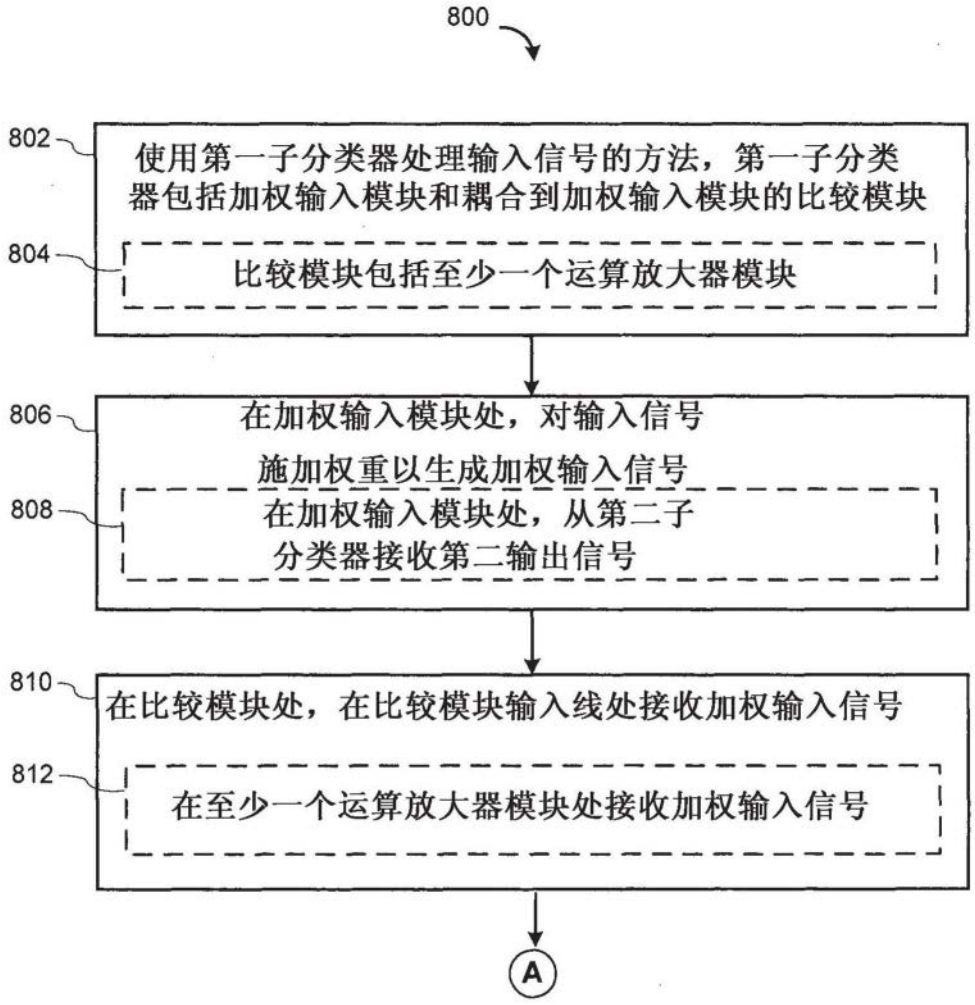

800
802
使用第一子分类器处理输入信号的方法，第一子分类器包括加权输入模块和耦合到加权输入模块的比较模块
804
比较模块包括至少一个运算放大器模块
806
在加权输入模块处，对输入信号施加权重以生成加权输入信号
808
在加权输入模块处，从第二子分类器接收第二输出信号
810
在比较模块处，在比较模块输入线处接收加权输入信号
812
在至少一个运算放大器模块处接收加权输入信号
A

图8A

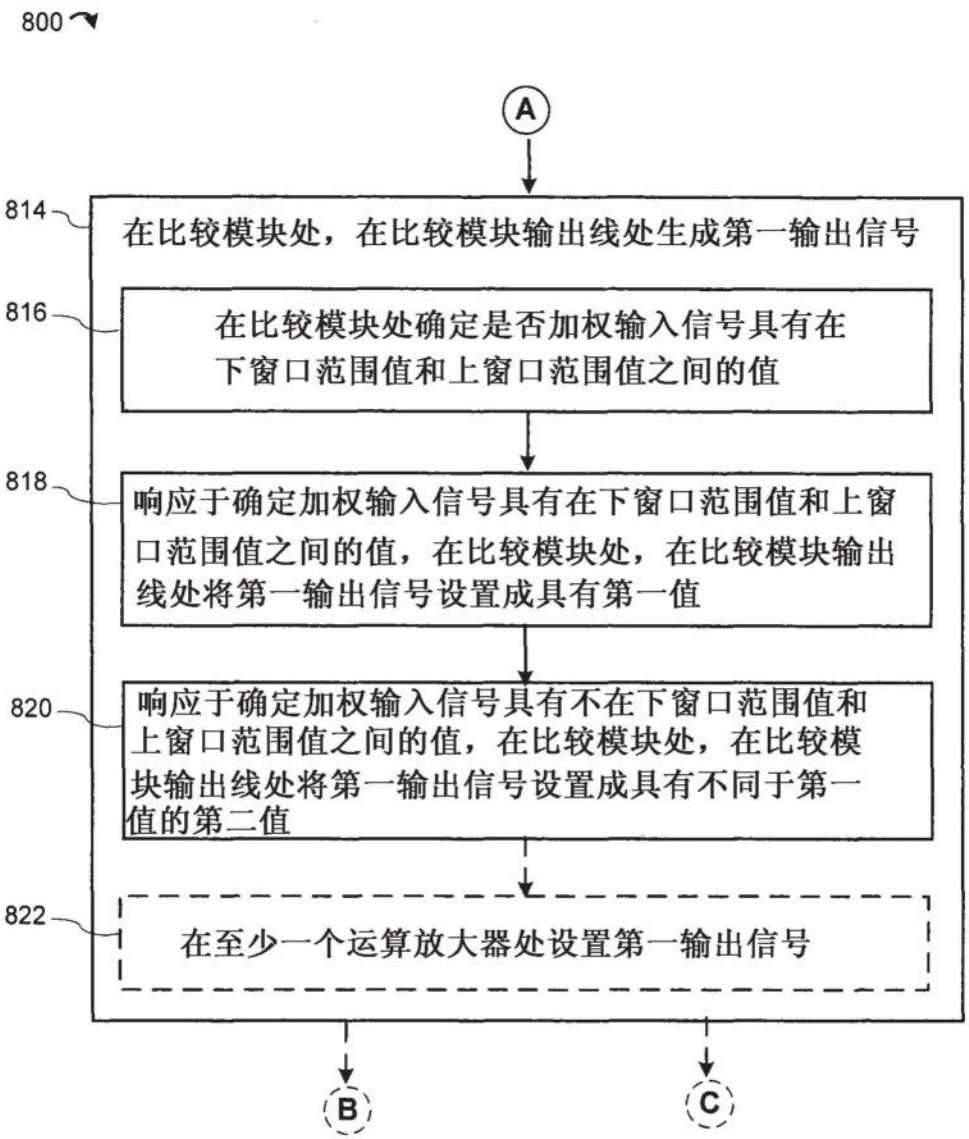

800
A
814
在比较模块处，在比较模块输出线处生成第一输出信号
816
在比较模块处确定是否加权输入信号具有在下窗口范围值和上窗口范围值之间的值
818
响应于确定加权输入信号具有在下窗口范围值和上窗口范围值之间的值，在比较模块处，在比较模块输出线处将第一输出信号设置成具有第一值
820
响应于确定加权输入信号具有不在下窗口范围值和上窗口范围值之间的值，在比较模块处，在比较模块输出线处将第一输出信号设置成具有不同于第一值的第二值
822
在至少一个运算放大器处设置第一输出信号
B
C

图8B

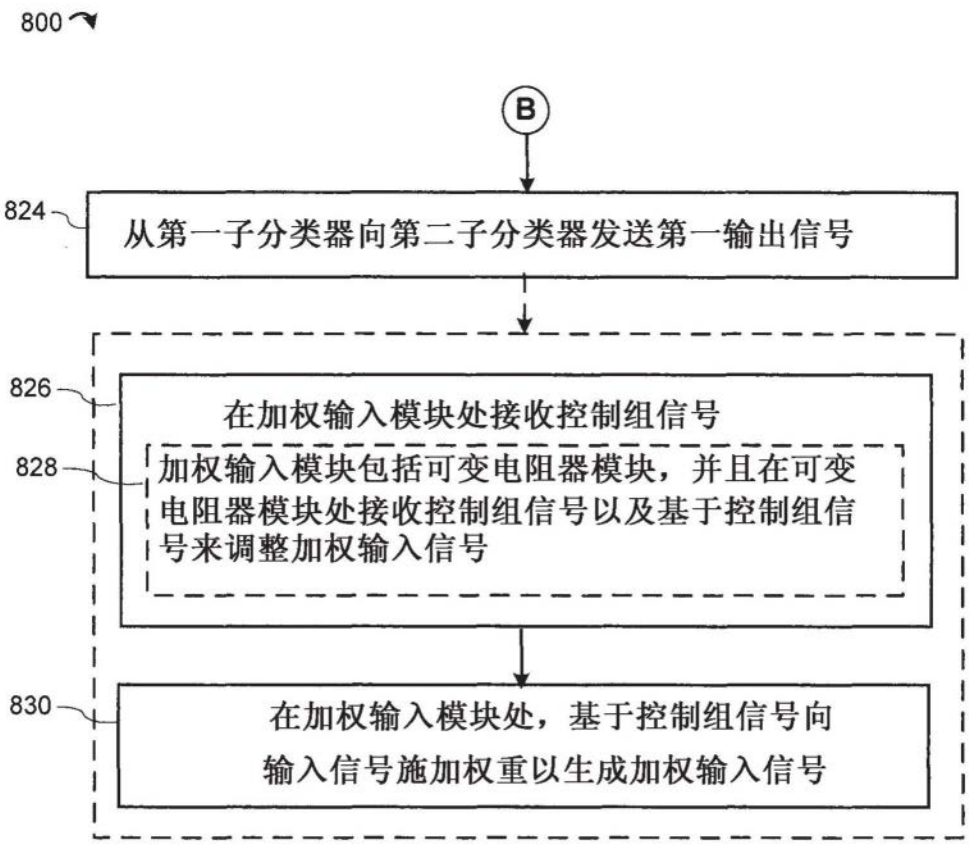

800
B
824
从第一子分类器向第二子分类器发送第一输出信号
826
在加权输入模块处接收控制组信号
828
加权输入模块包括可变电阻器模块，并且在可变电阻器模块处接收控制组信号以及基于控制组信号来调整加权输入信号
830
在加权输入模块处，基于控制组信号向输入信号施加权重以生成加权输入信号

图8C

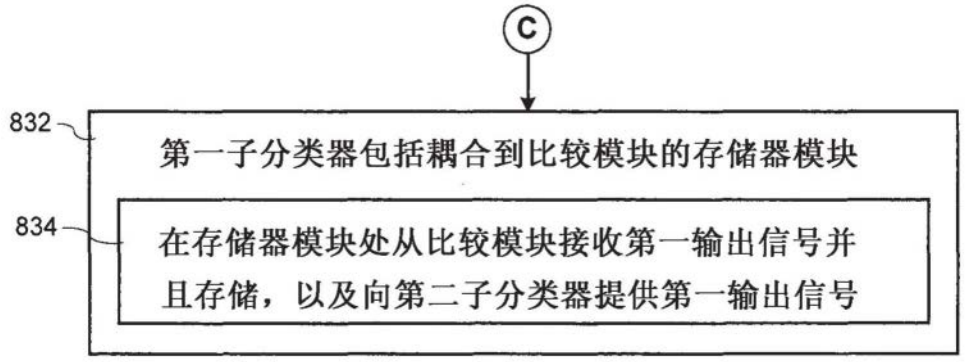

C
832
第一子分类器包括耦合到比较模块的存储器模块
834
在存储器模块处从比较模块接收第一输出信号并且存储，以及向第二子分类器提供第一输出信号

图8D

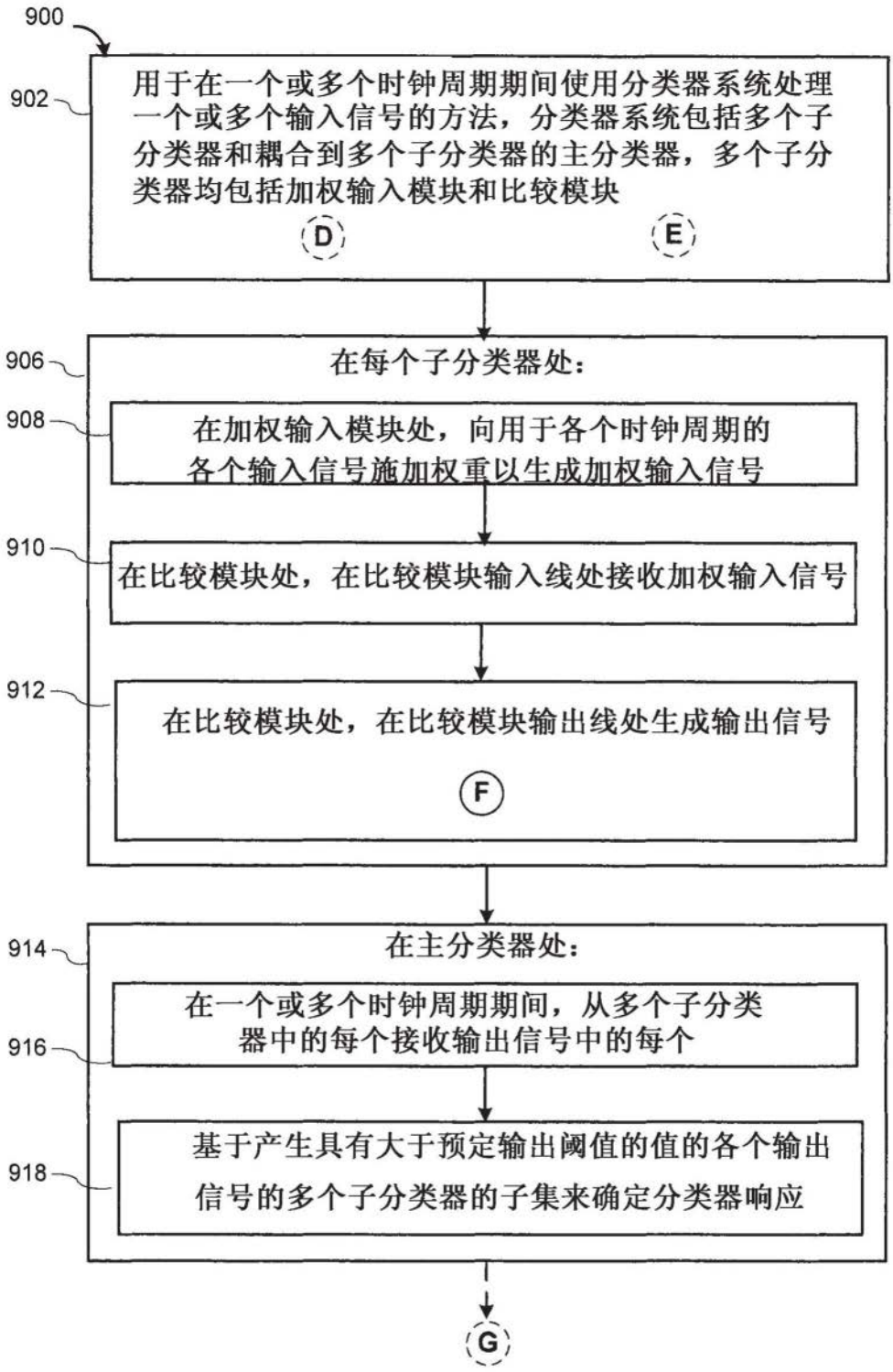

900
902
用于在一个或多个时钟周期期间使用分类器系统处理一个或多个输入信号的方法，分类器系统包括多个子分类器和耦合到多个子分类器的主分类器，多个子分类器均包括加权输入模块和比较模块
D
E
906
在每个子分类器处：
908
在加权输入模块处，向用于各个时钟周期的各个输入信号施加权重以生成加权输入信号
910
在比较模块处，在比较模块输入线处接收加权输入信号
912
在比较模块处，在比较模块输出线处生成输出信号
F
914
在主分类器处：
916
在一个或多个时钟周期期间，从多个子分类器中的每个接收输出信号中的每个
918
基于产生具有大于预定输出阈值的值的各个输出信号的多个子分类器的子集来确定分类器响应
G

图9A

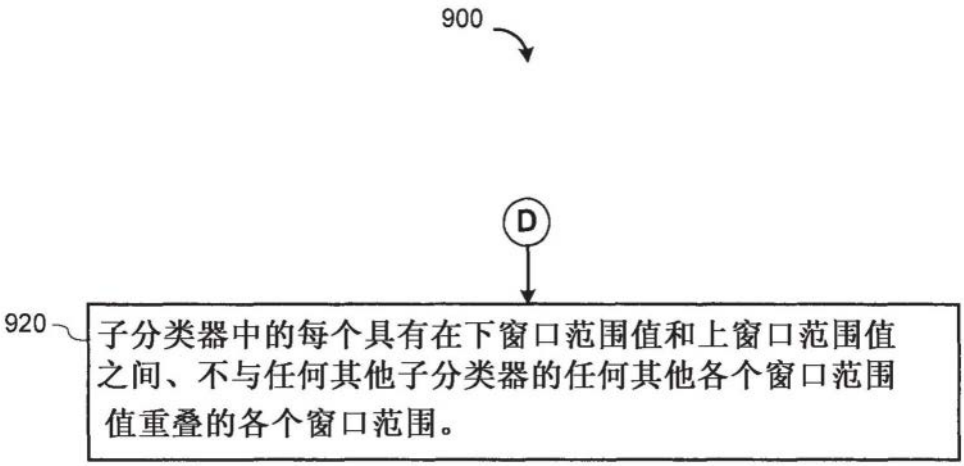

900
D
920
子分类器中的每个具有在下窗口范围值和上窗口范围值之间、不与任何其他子分类器的任何其他各个窗口范围值重叠的各个窗口范围。

图9B

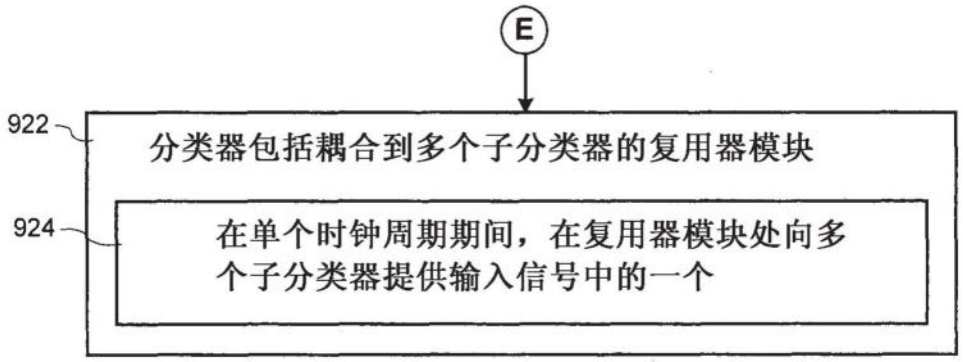

E
922
分类器包括耦合到多个子分类器的复用器模块
924
在单个时钟周期期间，在复用器模块处向多个子分类器提供输入信号中的一个

图9C

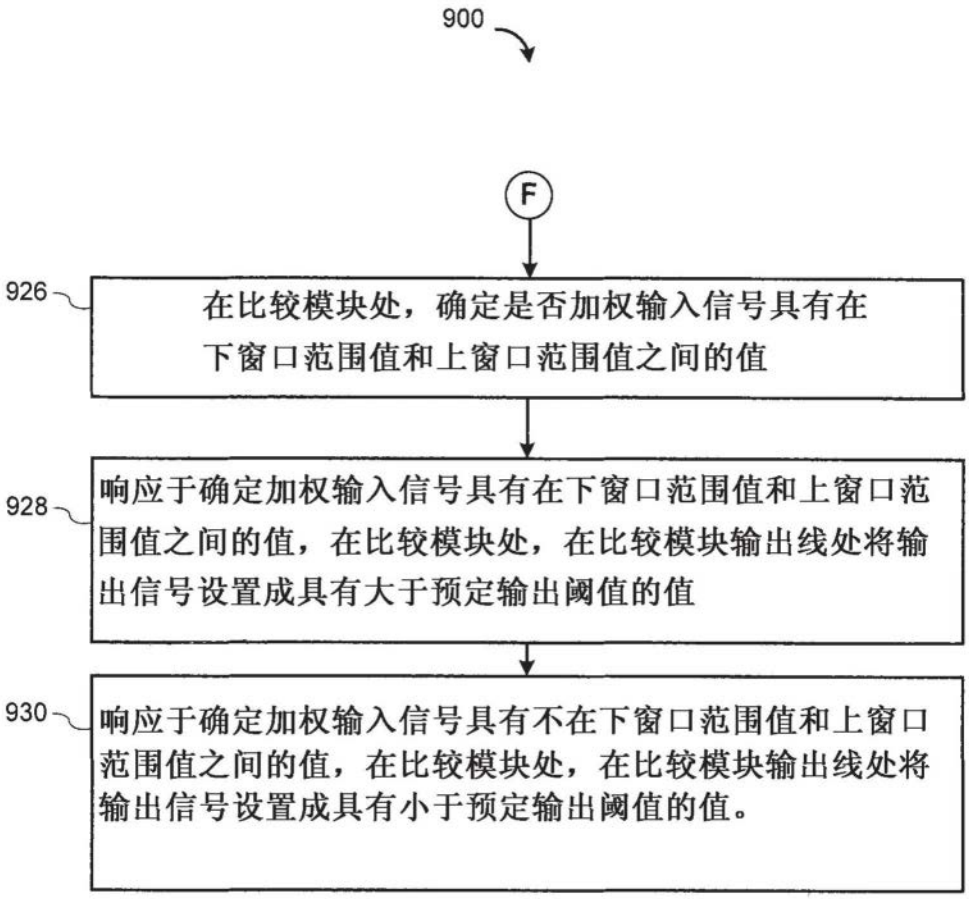

900
F
926
在比较模块处，确定是否加权输入信号具有在下窗口范围值和上窗口范围值之间的值
928
响应于确定加权输入信号具有在下窗口范围值和上窗口范围值之间的值，在比较模块处，在比较模块输出线处将输出信号设置成具有大于预定输出阈值的值
930
响应于确定加权输入信号具有不在下窗口范围值和上窗口范围值之间的值，在比较模块处，在比较模块输出线处将输出信号设置成具有小于预定输出阈值的值。

图9D

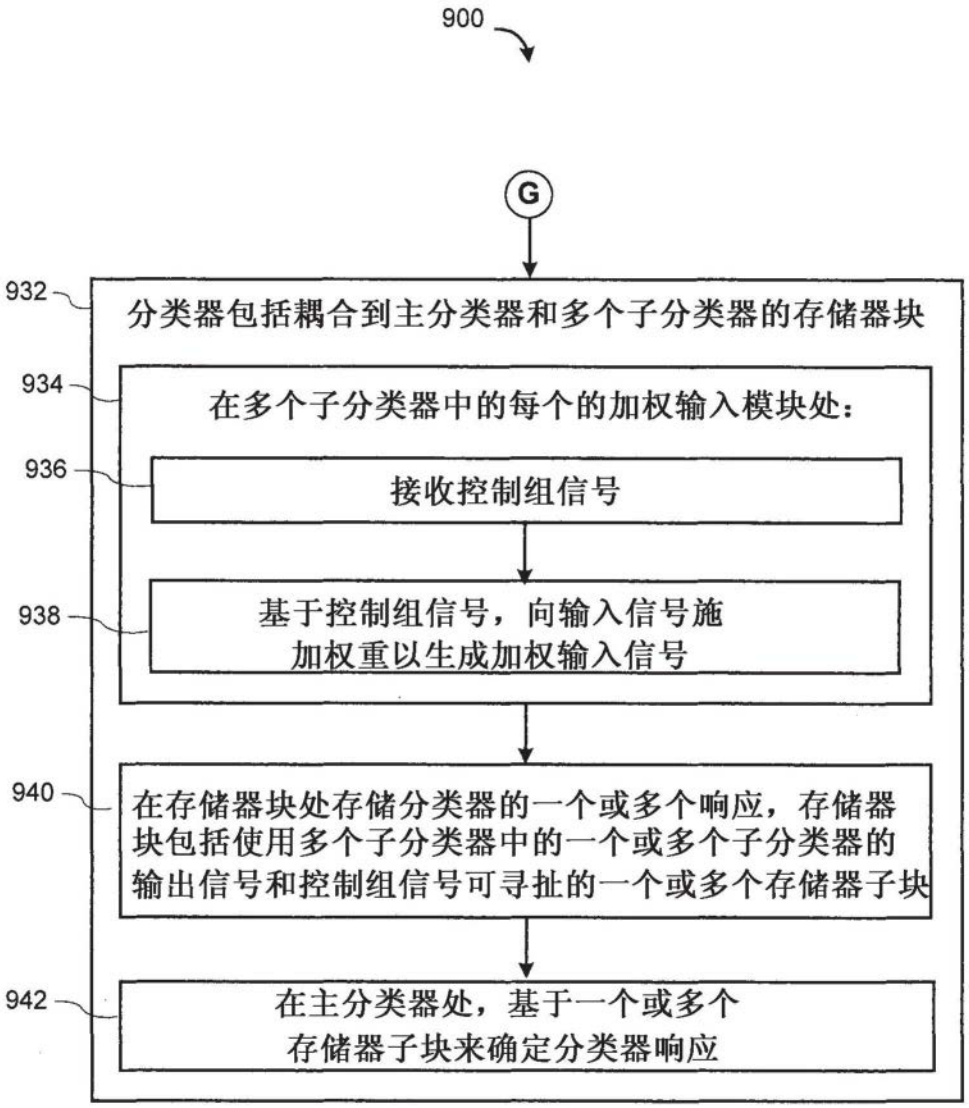

900
G
932
分类器包括耦合到主分类器和多个子分类器的存储器块
934
在多个子分类器中的每个的加权输入模块处：
936
接收控制组信号
938
基于控制组信号，向输入信号施加权重以生成加权输入信号
940
在存储器块处存储分类器的一个或多个响应，存储器块包括使用多个子分类器中的一个或多个子分类器的输出信号和控制组信号可寻扯的一个或多个存储器子块
942
在主分类器处，基于一个或多个存储器子块来确定分类器响应

图9E