(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 107077637 B
(45)授权公告日 2021.07.20

(21)申请号 201580015756.6

(22)申请日 2015.03.17

(65)同一申请的已公布的文献号
申请公布号 CN 107077637 A

(43)申请公布日 2017.08.18

(30)优先权数据
61/969,747 2014.03.24 US
14/513,155 2014.10.13 US

(85)PCT国际申请进入国家阶段日
2016.09.22

(86)PCT国际申请的申请数据
PCT/US2015/021077 2015.03.17

(87)PCT国际申请的公布数据
WO2015/148189 EN 2015.10.01

(73)专利权人 高通股份有限公司
地址 美国加利福尼亚州

(72)发明人 V·S·R·安纳普莱蒂
D·J·朱里安 R·B·托瓦
Y·刘

(74)专利代理机构 上海专利商标事务所有限公司 31100
代理人 李小芳

(51)Int.Cl.
*G06N 3/04*(2006.01)

(56)对比文件
CN 1622129 A,2005.06.01
CN 101310294 A,2008.11.19
Jerome T. Connor等.Recurrent Neural Networks and Robust Time Series Prediction.《IEEE TRANSACTIONS ON NEURAL NETWORKS》.1994,第5卷(第2期),第240-254页.
Stuart W. Perry等.A Recurrent Neural Network for Detecting Objects in Sequences of Sector-Scan Sonar Images.《IEEE JOURNAL OF OCEANIC ENGINEERING》.2004,第29卷(第3期),第857-871页.
Y.HUANG等.Predictive coding.《WILEY INTERDISCRIPLINARY REVIEWS:COGNITIVE SCIENCE》.2011,第2卷(第5期),第580-593页.
Jerome T. Connor等.Recurrent Neural Networks and Robust Time Series Prediction.《IEEE TRANSACTIONS ON NEURAL NETWORKS》.1994,第5卷(第2期),第240-254页.

审查员 张玲

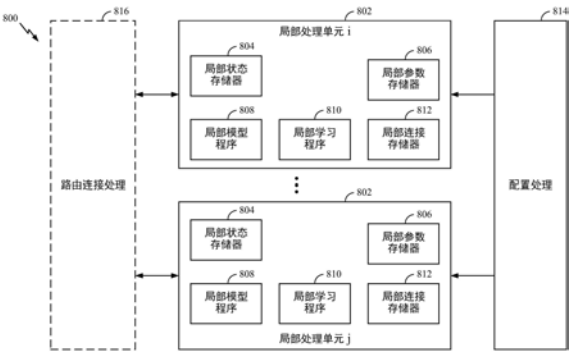
权利要求书2页 说明书16页 附图7页

(54)发明名称

神经网络中的差分编码

(57)摘要

神经网络中的差分编码包括基于神经网络中的神经元的至少一个先前激活值来预测该神经元的激活值。该编码进一步包括基于神经网络中的该神经元的预测激活值与实际激活值之间的差值来对值进行编码。




CN 107077637 B

1.一种用于在人工神经网络中执行针对图像的差分编码的方法,包括:

由所述神经网络的第一层的第一神经元基于所述第一神经元的激活历史以及提供给所述第一神经元的附加值来预测所述第一神经元的激活值,所述附加值是至少部分地基于图像运动估计来计算的;

由所述第一神经元至少部分地基于所述神经网络中的所述第一神经元的预测激活值与实际激活值之间的差值来对值进行编码;以及

将经编码的值传达给所述神经网络的第二层的第二神经元而不传播所述实际激活值。

2.如权利要求1所述的方法,其中,所传达的经编码值进一步基于所述预测激活值与所述实际激活值之间的取阈差值。

3.如权利要求1所述的方法,其中,所传达的经编码值是至少部分地基于所述经编码值的位数来选择的。

4.如权利要求1所述的方法,其中,所述预测是至少部分地基于接收到输入来执行的。

5.如权利要求1所述的方法,进一步包括至少部分地基于所述值的位宽来对所述值进行编码。

6.如权利要求1所述的方法,其中,所述编码是至少部分地基于以神经网络输出为基础的触发来执行的。

7.如权利要求1所述的方法,其中,所述编码相对于至所述神经网络的输入被延迟。

8.如权利要求1所述的方法,其中,所述编码进一步至少部分地基于所述神经网络的输出。

9.如权利要求1所述的方法,其中,当输入-输出关系是确定性的时,所述至少一个先前激活值包括输入历史。

10.如权利要求1所述的方法,其中,当输入-输出关系是随机的时,所述至少一个先前激活值包括输入历史和输出历史。

11.如权利要求1所述的方法,进一步包括至少部分地基于预测输入值来计算所述预测激活值。

12.如权利要求11所述的方法,进一步包括通过将所述经编码值与所述预测输入值相组合来计算实际输入值。

13.如权利要求11所述的方法,其中,计算所述预测输入值和所述预测激活值包括使用所述神经元的多个先前输入值和多个先前激活值的线性组合。

14.如权利要求1所述的方法,其中,所述神经元的所述预测激活值至少部分地基于所述神经元的状态以及至所述神经元的输入。

15.如权利要求14所述的方法,其中,所述神经元的状态是基于以下至少一者来更新的:先前状态、输入值、输出值、预测激活值、以及定向激活值。

16.如权利要求14所述的方法,其中,所述神经元的状态包括以下至少一者:输入历史、输出历史、预测激活值历史、以及定向激活值历史。

17.如权利要求16所述的方法,其中,所述预测至少部分地基于另一神经元的状态。

18.如权利要求1所述的方法,其中,所述预测激活值至少部分地基于多个先前实际激活值的线性组合或先前输入值的线性组合。

19.如权利要求1所述的方法,其中,所述附加值包括来自另一神经元的反馈信号。

20.一种用于在人工神经网络中执行针对图像的差分编码的装备,包括:

用于通过所述神经网络的第一层的第一神经元基于所述第一神经元的激活历史以及提供给所述第一神经元的附加值来预测所述第一神经元的激活值的装置,所述附加值是至少部分地基于图像运动估计来计算的;

用于通过所述第一神经元至少部分地基于所述神经网络中的所述第一神经元的预测激活值与实际激活值之间的差值来对值进行编码的装置;以及

用于将经编码的值传达给所述神经网络的第二层的第二神经元而不传播所述实际激活值的装置。

21.如权利要求20所述的装备,进一步包括用于至少部分地基于接收到输入来预测所述激活值的装置。

22.如权利要求20所述的装备,进一步包括用于至少部分地基于预测输入值来计算所述预测激活值的装置。

23.如权利要求22所述的装备,进一步包括用于通过将所述经编码值与所述预测输入值相组合来计算实际输入值的装置。

24.一种其上编码有用于在人工神经网络中执行针对图像的差分编码的程序代码的非瞬态计算机可读介质,所述程序代码由处理器执行并且包括:

用于通过所述神经网络的第一层的第一神经元基于所述第一神经元的激活历史以及提供给所述第一神经元的附加值来预测所述第一神经元的激活值的程序代码,所述附加值是至少部分地基于图像运动估计来计算的;

用于通过所述第一神经元至少部分地基于所述神经网络中的所述第一神经元的预测激活值与实际激活值之间的差值来对值进行编码的程序代码;以及

用于将经编码的值传达给所述神经网络的第二层的第二神经元而不传播所述实际激活值的程序。

25.如权利要求24所述的非瞬态计算机可读介质,进一步包括用于至少部分地基于接收到输入来预测所述激活值的程序代码。

26.如权利要求24所述的非瞬态计算机可读介质,进一步包括用于至少部分地基于预测输入值来计算所述预测激活值的程序代码。

27.如权利要求26所述的非瞬态计算机可读介质,进一步包括用于通过将所述经编码值与所述预测输入值相组合来计算实际输入值的程序代码。

# 神经网络中的差分编码

[0001] 相关申请的交叉引用

[0002] 本申请依据35U.S.C.§119(e)要求于2014年3月24日提交的题为“DIFFERENTIAL ENCODING IN NEURAL NETWORKS(神经网络中的差分编码)”的美国临时专利申请No.61/969,747的权益,其公开内容全部通过援引明确纳入于此。

[0003] 背景

[0004] 领域

[0005] 本公开的某些方面一般涉及神经系统工程,尤其涉及用于神经网络中的差分编码的系统和方法。

## 背景技术

[0006] 可包括一群互连的人工神经元(即,神经元模型)的人工神经网络是一种计算设备或者表示将由计算设备执行的方法。人工神经网络可具有生物学神经网络中的对应的结构和/或功能。然而,人工神经网络可为其中传统计算技术是麻烦的、不切实际的、或不胜任的某些应用提供创新且有用的计算技术。由于人工神经网络能从观察中推断出功能,因此这样的网络在因任务或数据的复杂度使得通过常规技术来设计该功能较为麻烦的应用中是特别有用的。

[0007] 概述

[0008] 根据本公开的一方面的一种在神经网络中执行差分编码的方法包括基于神经网络中的神经元的至少一个先前激活值来预测该神经元的激活值。此类方法进一步包括基于神经网络中的该神经元的预测激活值与激活值之间的差值来对值进行编码。

[0009] 根据本公开的一方面的一种用于在神经网络中执行差分编码的装置包括存储器和耦合到该存储器的至少一个处理器。该(诸)处理器被配置成基于神经网络中的神经元的至少一个先前激活值来预测该神经元的激活值。该(诸)处理器还被配置成基于神经网络中的该神经元的预测激活值与激活值之间的差值来对值进行编码。

[0010] 根据本公开的另一方面的一种用于在尖峰神经网络中执行差分编码的装备包括用于基于神经网络中的神经元的至少一个先前激活值来预测该神经元的激活值的装置。此类装备进一步包括用于基于神经网络中的该神经元的预测激活值与激活值之间的差值来对值进行编码的装置。

[0011] 根据本公开另一方面的一种用于在尖峰神经网络中执行差分编码的计算机程序产品包括其上编码有程序代码的非瞬态计算机可读介质。该程序代码包括用于基于神经网络中的神经元的至少一个先前激活值来预测该神经元的激活值的程序代码。该程序代码还包括用于基于神经网络中的该神经元的预测激活值与激活值之间的差值来对值进行编码的程序代码。

[0012] 这已较宽泛地勾勒出本公开的特征和技术优势以便下面的详细描述可以被更好地理解。本公开的附加特征和优点将在下文描述。本领域技术人员应该领会,本公开可容易地被用作修改或设计用于实施与本公开相同的目的的其他结构的基础。本领域技术人员还

应认识到，这样的等效构造并不脱离所附权利要求中所阐述的本公开的教导。被认为是本公开的特性的新颖特征在其组织和操作方法两方面连同进一步的目的和优点在结合附图来考虑以下描述时将被更好地理解。然而，要清楚理解的是，提供每一幅附图均仅用于解说和描述目的，且无意作为对本公开的限定的定义。

[0013] 附图简述

[0014] 在结合附图理解下面阐述的详细描述时，本公开的特征、本质和优点将变得更加明显，在附图中，相同附图标记始终作相应标识。

[0015] 图1解说了根据本公开的某些方面的示例神经元网络。

[0016] 图2解说了根据本公开的某些方面的计算网络(神经系统或神经网络)的处理单元(神经元)的示例。

[0017] 图3解说了根据本公开的某些方面的尖峰定时依赖可塑性(STDP)曲线的示例。

[0018] 图4解说了根据本公开的某些方面的用于定义神经元模型的行为的正态相和负态相的示例。

[0019] 图5解说了根据本公开的某些方面的使用通用处理器来设计神经网络的示例实现。

[0020] 图6解说了根据本公开的某些方面的设计其中存储器可以与个体分布式处理单元对接的神经网络的示例实现。

[0021] 图7解说了根据本公开的某些方面的基于分布式存储器和分布式处理单元来设计神经网络的示例实现。

[0022] 图8解说了根据本公开的某些方面的神经网络的示例实现。

[0023] 图9解说了根据本公开的各方面的用于执行差分编码的方法。

[0024] 详细描述

[0025] 以下结合附图阐述的详细描述旨在作为各种配置的描述，而无意表示可实践本文中所描述的概念的仅有的配置。本详细描述包括具体细节以便提供对各种概念的透彻理解。然而，对于本领域技术人员将显而易见的是，没有这些具体细节也可实践这些概念。在一些实例中，以框图形式示出众所周知的结构和组件以避免湮没此类概念。

[0026] 基于本教导，本领域技术人员应领会，本公开的范围旨在覆盖本公开的任何方面，不论其是与本公开的任何其他方面相独立地还是组合地实现的。例如，可以使用所阐述的任何数目的方面来实现装置或实践方法。另外，本公开的范围旨在覆盖使用作为所阐述的本公开的各个方面的补充或者与之不同的其他结构、功能性、或者结构及功能性来实践的此类装置或方法。应当理解，所披露的本公开的任何方面可由权利要求的一个或多个元素来实施。

[0027] 措辞“示例性”在本文中用于表示“用作示例、实例或解说”。本文中描述为“示例性”的任何方面不必被解释为优于或胜过其他方面。

[0028] 尽管本文描述了特定方面，但这些方面的众多变体和置换落在本公开的范围之内。虽然提到了优选方面的一些益处和优点，但本公开的范围并非旨在被限定于特定益处、用途或目标。相反，本公开的各方面旨在能宽泛地应用于不同的技术、系统配置、网络和协议，其中一些作为示例在附图以及以下对优选方面的描述中解说。详细描述和附图仅仅解说本公开而非限定本公开，本公开的范围由所附权利要求及其等效技术方案来定义。

[0029] 示例神经系统、训练及操作

[0030] 图1解说了根据本公开的某些方面的具有多级神经元的示例人工神经系统100。神经系统100可具有神经元级102，该神经元级102通过突触连接网络104(即，前馈连接)来连接到另一神经元级106。为简单起见，图1中仅解说了两级神经元，尽管神经系统中可存在更少或更多级神经元。应注意，一些神经元可通过侧向连接来连接至同层中的其他神经元。此外，一些神经元可通过反馈连接来后向连接至先前层中的神经元。

[0031] 如图1所解说的，级102中的每一个神经元可以接收可由前级的神经元(未在图1中示出)生成的输入信号108。信号108可表示级102的神经元的输入电流。该电流可在神经元膜上累积以对膜电位进行充电。当膜电位达到其阈值时，该神经元可激发并生成输出尖峰，该输出尖峰将被传递到下一级神经元(例如，级106)。在一些建模办法中，神经元可以连续地向下一级神经元传递信号。该信号通常是膜电位的函数。此类行为可在硬件和/或软件(包括模拟和数字实现，诸如以下所述那些实现)中进行仿真或模拟。

[0032] 在生物学神经元中，在神经元激发时生成的输出尖峰被称为动作电位。该电信号是相对迅速、瞬态的神经脉冲，其具有约为100mV的振幅和约为1ms的历时。在具有一系列连通的神经元(例如，尖峰从图1中的一级神经元传递至另一级神经元)的神经系统的特定实施例中，每个动作电位都具有基本上相同的振幅和历时，并且因此该信号中的信息可仅由尖峰的频率和数目、或尖峰的时间来表示，而不由振幅来表示。动作电位所携带的信息可由尖峰、发放了尖峰的神经元、以及该尖峰相对于一个或数个其他尖峰的时间来确定。尖峰的重要性可由向各神经元之间的连接所应用的权重来确定，如以下所解释的。

[0033] 尖峰从一级神经元向另一级神经元的传递可通过突触连接(或简称“突触”)网络104来达成，如图1中所解说的。相对于突触104，级102的神经元可被视为突触前神经元，而级106的神经元可被视为突触后神经元。突触104可接收来自级102的神经元的输出信号(即，尖峰)，并根据可调节突触权重 $w_1^{(i,i+1)}$、…、$w_P^{(i,i+1)}$ 来按比例缩放那些信号，其中P是级102的神经元与级106的神经元之间的突触连接的总数，并且i是神经元级的指示符。在图1的示例中，i表示神经元级102并且i+1表示神经元级106。此外，经按比例缩放的信号可被组合以作为级106中每个神经元的输入信号。级106中的每个神经元可基于对应的组合输入信号来生成输出尖峰110。可使用另一突触连接网络(图1中未示出)将这些输出尖峰110传递到另一级神经元。

[0034] 生物学突触可以仲裁突触后神经元中的兴奋性或抑制性(超极化)动作，并且还可用于放大神经元信号。兴奋性信号使膜电位去极化(即，相对于静息电位增大膜电位)。如果在某个时间段内接收到足够的兴奋性信号以使膜电位去极化到高于阈值，则在突触后神经元中发生动作电位。相反，抑制性信号一般使膜电位超极化(即，降低膜电位)。抑制性信号如果足够强则可抵消掉兴奋性信号之和并阻止膜电位到达阈值。除了抵消掉突触兴奋以外，突触抑制还可对自发活跃神经元施加强力的控制。自发活跃神经元是指在没有进一步输入的情况下(例如，由于其动态或反馈而)发放尖峰的神经元。通过压制这些神经元中的动作电位的自发生成，突触抑制可对神经元中的激发模式进行定形，这一般被称为雕刻。取决于期望的行为，各种突触104可充当兴奋性或抑制性突触的任何组合。

[0035] 神经系统100可由通用处理器、数字信号处理器(DSP)、专用集成电路(ASIC)、现场

可编程门阵列(FPGA)或其他可编程逻辑器件(PLD)、分立的门或晶体管逻辑、分立的硬件组件、由处理器执行的软件模块、或其任何组合来仿真。神经系统100可用在大范围的应用中，诸如图像和模式识别、机器学习、电机控制、及类似应用等。神经系统100中的每一神经元可被实现为神经元电路。被充电至发起输出尖峰的阈值的神经元膜可被实现为例如对流经其的电流进行积分的电容器。

[0036] 在一方面，电容器作为神经元电路的电流积分器件可被除去，并且可使用较小的忆阻器元件来替代它。这种办法可应用于神经元电路中，以及其中大容量电容器被用作电流积分器的各种其他应用中。另外，每个突触104可基于忆阻器元件来实现，其中突触权重变化可与忆阻器电阻的变化有关。使用纳米特征尺寸的忆阻器，可显著地减小神经元电路和突触的面积，这可使得实现大规模神经系统硬件实现更为切实可行。

[0037] 对神经系统100进行仿真的神经处理器的功能性可取决于突触连接的权重，这些权重可控制神经元之间的连接的强度。突触权重可存储在非易失性存储器中以在掉电之后保留该处理器的功能性。在一方面，突触权重存储器可实现在与主神经处理器芯片分开的外部芯片上。突触权重存储器可与神经处理器芯片分开地封装成可更换的存储卡。这可向神经处理器提供多种多样的功能性，其中特定功能性可基于当前附连至神经处理器的存储卡中所存储的突触权重。

[0038] 图2解说了根据本公开的某些方面的计算网络(例如，神经系统或神经网络)的处理单元(例如，神经元或神经元电路)202的示例性示图200。例如，神经元202可对应于来自图1的级102和106的任何神经元。神经元202可接收多个输入信号$204_1$-$204_N$，这些输入信号可以是该神经系统外部的信号、或是由同一神经系统的其他神经元所生成的信号、或这两者。输入信号可以是电流、电导、电压、实数值的和/或复数值的。输入信号可包括具有定点或浮点表示的数值。可通过突触连接将这些输入信号递送到神经元202，突触连接根据可调节突触权重$206_1$-$206_N$($W_1$-$W_N$)对这些信号进行按比例缩放，其中N可以是神经元202的输入连接总数。

[0039] 神经元202可组合这些经按比例缩放的输入信号，并且使用组合的经按比例缩放的输入来生成输出信号208(即，信号Y)。输出信号208可以是电流、电导、电压、实数值的和/或复数值的。输出信号可以是具有定点或浮点表示的数值。随后该输出信号208可作为输入信号传递至同一神经系统的其他神经元、或作为输入信号传递至同一神经元202、或作为该神经系统的输出来传递。

[0040] 处理单元(神经元)202可由电路来仿真，并且其输入和输出连接可由具有突触电路的电连接来仿真。处理单元202及其输入和输出连接也可由软件代码来仿真。处理单元202还可由电路来仿真，而其输入和输出连接可由软件代码来仿真。在一方面，计算网络中的处理单元202可以是模拟电路。在另一方面，处理单元202可以是数字电路。在又一方面，处理单元202可以是具有模拟和数字组件两者的混合信号电路。计算网络可包括任何前述形式的处理单元。使用这样的处理单元的计算网络(神经系统或神经网络)可用在大范围的应用中，诸如图像和模式识别、机器学习、电机控制、及类似应用等。

[0041] 在神经网络的训练过程期间，突触权重(例如，来自图1的权重$w_1^{(i,i+1)}$、...、$w_P^{(i,i+1)}$和/或来自图2的权重$206_1$-$206_N$)可用随机值来初始化并根据学习规则而被增大或减小。本

领域技术人员将领会，学习规则的示例包括但不限于尖峰定时依赖可塑性(STDP)学习规则、Hebb规则、Oja规则、Bienenstock-Copper-Munro(BCM)规则等。在一些方面，这些权重可稳定或收敛至两个值(即，权重的双峰分布)之一。该效应可被用于减少每个突触权重的位数、提高从/向存储突触权重的存储器读取和写入的速度、以及降低突触存储器的功率和/或处理器消耗。

[0042] 突触类型

[0043] 在神经网络的硬件和软件模型中，突触相关功能的处理可基于突触类型。突触类型可以是非可塑突触(权重和延迟没有改变)、可塑突触(权重可改变)、结构化延迟可塑突触(权重和延迟可改变)、全可塑突触(权重、延迟和连通性可改变)、以及基于此的变型(例如，延迟可改变，但在权重或连通性方面没有改变)。多种类型的优点在于处理可以被细分。例如，非可塑突触不会使用待执行的可塑性功能(或等待此类功能完成)。类似地，延迟和权重可塑性可被细分成可一起或分开地、顺序地或并行地运作的操作。不同类型的突触对于适用的每一种不同的可塑性类型可具有不同的查找表或公式以及参数。因此，这些方法将针对该突触的类型来访问相关的表、公式或参数。

[0044] 还进一步牵涉到以下事实：尖峰定时依赖型结构化可塑性可独立于突触可塑性地来执行。结构化可塑性即使在权重幅值没有改变的情况下(例如，如果权重已达最小或最大值、或者其由于某种其他原因而不被改变)也可被执行，因为结构化可塑性(即，延迟改变的量)可以是pre-post(前-后)尖峰时间差的直接函数。替换地，结构化可塑性可被设为权重变化量的函数或者可基于与权重或权重变化的界限有关的条件来设置。例如，突触延迟可仅在权重变化发生时或者在权重到达0的情况下才改变，但在这些权重为最大值时则不改变。然而，具有独立函数以使得这些过程能被并行化从而减少存储器访问的次数和交叠可能是有利的。

[0045] 突触可塑性的确定

[0046] 神经元可塑性(或简称“可塑性”)是大脑中的神经元和神经网络响应于新的信息、感官刺激、发展、损坏、或机能障碍而改变其突触连接和行为的能力。可塑性对于生物学中的学习和记忆、以及对于计算神经元科学和神经网络是重要的。已经研究了各种形式的可塑性，诸如突触可塑性(例如，根据Hebbian理论)、尖峰定时依赖可塑性(STDP)、非突触可塑性、活跃性依赖可塑性、结构化可塑性和自稳态可塑性。

[0047] STDP是调节神经元之间的突触连接的强度的学习过程。连接强度是基于特定神经元的输出与收到输入尖峰(即，动作电位)的相对定时来调节的。在STDP过程下，如果至某个神经元的输入尖峰平均而言倾向于紧挨在该神经元的输出尖峰之前发生，则可发生长期增强(LTP)。于是使得该特定输入在一定程度上更强。另一方面，如果输入尖峰平均而言倾向于紧接在输出尖峰之后发生，则可发生长期抑压(LTD)。于是使得该特定输入在一定程度上更弱，并由此得名“尖峰定时依赖可塑性”。因此，使得可能是突触后神经元兴奋原因的输入甚至在将来作出贡献的可能性更大，而使得不是突触后尖峰的原因的输入在将来作出贡献的可能性更小。该过程继续，直至初始连接集合的子集保留，而所有其他连接的影响减小至无关紧要的水平。

[0048] 由于神经元一般在其许多输入都在一短时段内发生(即，累积性足以引起输出)时产生输出尖峰，因此通常保留下来的输入子集包括倾向于在时间上相关的那些输入。另外，

由于在输出尖峰之前发生的输入被加强，因此提供对相关性的最早充分累积性指示的那些输入将最终变成至该神经元的最后输入。

[0049] STDP学习规则可因变于突触前神经元的尖峰时间$t_{pre}$与突触后神经元的尖峰时间$t_{post}$之间的时间差(即，$t = t_{post} - t_{pre}$)来有效地适配将该突触前神经元连接到该突触后神经元的突触的突触权重。STDP的典型公式化是若该时间差为正(突触前神经元在突触后神经元之前激发)则增大突触权重(即，增强该突触)，以及若该时间差为负(突触后神经元在突触前神经元之前激发)则减小突触权重(即，抑压该突触)。

[0050] 在STDP过程中，突触权重随时间推移的改变可通常使用指数式衰退来达成，如由下式给出的：

[0051] $$\Delta w(t) = \begin{cases} a_+ e^{-t/k_+} + \mu, t > 0 \\ a_- e^{t/k_-}, t < 0 \end{cases}, \qquad (1)$$

[0052] 其中$k_+$和$k_-\tau sign(\Delta t)$分别是针对正和负时间差的时间常数，$a_+$和$a_-$是对应的比例缩放幅值，并且μ是可应用于正时间差和/或负时间差的偏移。

[0053] 图3解说了根据STDP，突触权重作为突触前(presynaptic)和突触后(postsynaptic)尖峰的相对定时的函数而改变的示例性示图300。如果突触前神经元在突触后神经元之前激发，则对应的突触权重可被增大，如曲线图300的部分302中所解说的。该权重增大可被称为该突触的LTP。从曲线图部分302可观察到，LTP的量可因变于突触前和突触后尖峰时间之差而大致呈指数式地下降。相反的激发次序可减小突触权重，如曲线图300的部分304中所解说的，从而导致该突触的LTD。

[0054] 如图3中的曲线图300中所解说的，可向STDP曲线图的LTP(因果性)部分302应用负偏移μ。x轴的交越点306(y＝0)可被配置成与最大时间滞后重合以考虑到来自层i-1的各因果性输入的相关性。在基于帧的输入(即，呈特定历时的包括尖峰或脉冲的帧的形式的输入)的情形中，可计算偏移值μ以反映帧边界。该帧中的第一输入尖峰(脉冲)可被视为要么如直接由突触后电位所建模地随时间衰退，要么在对神经状态的影响的意义上随时间衰退。如果该帧中的第二输入尖峰(脉冲)被视为与特定时间帧相关或有关，则该帧之前和之后的有关时间可通过使STDP曲线的一个或多个部分偏移以使得这些有关时间中的值可以不同(例如，对于大于一个帧为负，而对于小于一个帧为正)来在该时间帧边界处被分开并在可塑性意义上被不同地对待。例如，负偏移μ可被设为偏移LTP以使得曲线实际上在大于帧时间的pre-post时间处变得低于零并且它由此为LTD而非LTP的一部分。

[0055] 神经元模型及操作

[0056] 存在一些用于设计有用的尖峰发放神经元模型的一般原理。良好的神经元模型在以下两个计算态相(regime)方面可具有丰富的潜在行为：重合性检测和功能性计算。此外，良好的神经元模型应当具有允许时间编码的两个要素：输入的抵达时间影响输出时间，以及重合性检测能具有窄时间窗。最后，为了在计算上是有吸引力的，良好的神经元模型在连续时间上可具有闭合形式解，并且具有稳定的行为，包括在靠近吸引子和鞍点之处。换言之，有用的神经元模型是可实践且可被用于建模丰富的、现实的且生物学一致的行为并且可被用于对神经电路进行工程设计和反向工程两者的神经元模型。

[0057] 神经元模型可取决于事件，诸如输入抵达、输出尖峰或其他事件，无论这些事件是

内部的还是外部的。为了达成丰富的行为库，能展现复杂行为的状态机可能是期望的。如果事件本身的发生在撇开输入贡献(若有)的情况下能影响状态机并约束该事件之后的动态，则该系统的将来状态并非仅是状态和输入的函数，而是状态、事件和输入的函数。

[0058] 在一方面，神经元n可被建模为尖峰带漏泄积分激发神经元，其膜电压$v_n(t)$由以下动态来支配：

[0059] $$\frac{dv_n(t)}{dt}=\alpha v_n(t)+\beta\sum_m w_{m,n}y_m(t-\Delta t_{m,n}), \quad (2)$$

[0060] 其中α和β是参数，$w_{m,n}$是将突触前神经元m连接至突触后神经元n的突触的突触权重，以及$y_m(t)$是神经元m的尖峰发放输出，其可根据$\Delta t_{m,n}$被延迟达树突或轴突延迟才抵达神经元n的胞体。

[0061] 应注意，从建立了对突触后神经元的充分输入的时间直至该突触后神经元实际上激发的时间存在延迟。在动态尖峰发放神经元模型(诸如Izhikevich简单模型)中，如果在去极化阈值$v_t$与峰值尖峰电压$v_{peak}$之间有差量，则可引发时间延迟。例如，在该简单模型中，神经元胞体动态可由关于电压和恢复的微分方程对来支配，即：

[0062] $$\frac{dv}{dt}=(k(v-v_t)(v-v_r)-u+I)/C, \quad (3)$$

[0063] $$\frac{du}{dt}=a(b(v-v_r)-u), \quad (4)$$

[0064] 其中v是膜电位，u是膜恢复变量，k是描述膜电位v的时间尺度的参数，a是描述恢复变量u的时间尺度的参数，b是描述恢复变量u对膜电位v的阈下波动的敏感度的参数，$v_r$是膜静息电位，I是突触电流，以及C是膜的电容。根据该模型，神经元被定义为在$v>v_{peak}$时发放尖峰。

[0065] Hunzinger Cold模型

[0066] Hunzinger Cold神经元模型是能再现丰富多样的各种神经行为的最小双态相尖峰发放线性动态模型。该模型的一维或二维线性动态可具有两个态相，其中时间常数(以及耦合)可取决于态相。在阈下态相中，时间常数(按照惯例为负)表示漏泄通道动态，其一般作用于以生物学一致的线性方式使细胞返回到静息。阈上态相中的时间常数(按照惯例为正)反映抗漏泄通道动态，其一般驱动细胞发放尖峰，而同时在尖峰生成中引发等待时间。

[0067] 如图4中所解说的，该模型400的动态可被划分成两个(或更多个)态相。这些态相可被称为负态相402(也可互换地称为带漏泄积分激发(LIF)态相，勿与LIF神经元模型混淆)以及正态相404(也可互换地称为抗漏泄积分激发(ALIF)态相，勿与ALIF神经元模型混淆)。在负态相402中，状态在将来事件的时间趋向于静息($v_-$)。在该负态相中，该模型一般展现出时间输入检测性质及其他阈下行为。在正态相404中，状态趋向于尖峰发放事件($v_s$)。在该正态相中，该模型展现出计算性质，诸如取决于后续输入事件而引发发放尖峰的等待时间。在事件方面对动态进行公式化以及将动态分成这两个态相是该模型的基础特性。

[0068] 线性双态相二维动态(对于状态v和u)可按照惯例定义为：

[0069] $$\tau_\rho \frac{dv}{dt} = v + q_\rho \qquad (5)$$

[0070] $$-\tau_u \frac{du}{dt} = u + r \quad , \qquad (6)$$

[0071] 其中$q_\rho$和r是用于耦合的线性变换变量。

[0072] 符号ρ在本文中用于标示动态态相，在讨论或表达具体态相的关系时，按照惯例对于负态相和正态相分别用符号“-”或“+”来替换符号ρ。

[0073] 模型状态由膜电位(电压)v和恢复电流u来定义。在基本形式中，态相在本质上是由模型状态来决定的。该精确和通用的定义存在一些细微却重要的方面，但目前考虑该模型在电压v高于阈值($v_+$)的情况下处于正态相404中，否则处于负态相402中。

[0074] 态相相关时间常数包括负态相时间常数$\tau_-$和正态相时间常数$\tau_+$。恢复电流时间常数$\tau_u$通常是与态相无关的。出于方便起见，负态相时间常数$\tau_-$通常被指定为反映衰退的负量，从而用于电压演变的相同表达式可用于正态相，在正态相中指数和$\tau_+$将一般为正，正如$\tau_u$那样。

[0075] 这两个状态元素的动态可在发生事件之际通过使状态偏离其零倾线(null-cline)的变换来耦合，其中变换变量为：

[0076] $q_\rho = -\tau_\rho \beta u - v_\rho$ (7)

[0077] $r = \delta(v+\varepsilon)$， (8)

[0078] 其中δ、ε、β和$v_-$、$v_+$是参数。$v_\rho$的两个值是这两个态相的参考电压的基数。参数$v_-$是负态相的基电压，并且膜电位在负态相中一般将朝向$v_-$衰退。参数$v_+$是正态相的基电压，并且膜电位在正态相中一般将趋向于背离$v_+$。

[0079] v和u的零倾线分别由变换变量$q_\rho$和r的负数给出。参数δ是控制u零倾线的斜率的比例缩放因子。参数ε通常被设为等于$-v_-$。参数β是控制这两个态相中的v零倾线的斜率的电阻值。$\tau_\rho$时间常数参数不仅控制指数式衰退，还单独地控制每个态相中的零倾线斜率。

[0080] 该模型可被定义为在电压v达到值$v_S$时发放尖峰。随后，状态可在发生复位事件(其可以与尖峰事件完全相同)之际被复位：

[0081] $$v = \hat{v}_- \qquad (9)$$

[0082] $u = u + \Delta u$， (10)

[0083] 其中$\hat{v}_-$和Δu是参数。复位电压$\hat{v}_-$通常被设为$v_-$。

[0084] 依照瞬时耦合的原理，闭合形式解不仅对于状态是可能的(且具有单个指数项)，而且对于到达特定状态的时间也是可能的。闭合形式状态解为：

[0085] $$v(t+\Delta t) = \left(v(t) + q_\rho\right)e^{\frac{\Delta t}{\tau_\rho}} - q_\rho \qquad (11)$$

[0086] $$u(t+\Delta t) = \left(u(t) + r\right)e^{-\frac{\Delta t}{\tau_u}} - r \quad 。 \qquad (12)$$

[0087] 因此，模型状态可仅在发生事件之际被更新，诸如在输入(突触前尖峰)或输出(突触后尖峰)之际被更新。还可在任何特定时间(无论是否有输入或输出)执行操作。

[0088] 而且，依照瞬时耦合原理，突触后尖峰的时间可被预计，因此到达特定状态的时间

可提前被确定而无需迭代技术或数值方法(例如,欧拉数值方法)。给定了先前电压状态$v_0$,直至到达电压状态$v_f$之前的时间延迟由下式给出:

[0089] $$\Delta t = \tau_\rho \log \frac{v_f + q_\rho}{v_0 + q_\rho} 。 \quad (13)$$

[0090] 如果尖峰被定义为发生在电压状态v到达$v_S$的时间,则从电压处于给定状态v的时间起测量的直至发生尖峰前的时间量或即相对延迟的闭合形式解为:

[0091] $$\Delta t_s = \begin{cases} \tau_+ \log \frac{v_s + q_+}{v + q_+}, & \text{如果} v > \hat{v}_+ \\ \infty, & \text{否则} \end{cases} , \quad (14)$$

[0092] 其中$\hat{v}_+$通常被设为参数$v_+$,但其他变型可以是可能的。

[0093] 模型动态的以上定义取决于该模型是在正态相还是负态相中。如所提及的,耦合和态相ρ可基于事件来计算。出于状态传播的目的,态相和耦合(变换)变量可基于在上一(先前)事件的时间的状态来定义。出于随后预计尖峰输出时间的目的,态相和耦合变量可基于在下一(当前)事件的时间的状态来定义。

[0094] 存在对该Cold模型、以及在时间上执行模拟、仿真、或建模的若干可能实现。这包括例如事件-更新、步阶-事件更新、以及步阶-更新模式。事件更新是其中基于事件或"事件更新"(在特定时刻)来更新状态的更新。步点更新是以间隔(例如,1ms)来更新模型的情况下的更新。这不一定利用迭代方法或数值方法。通过仅在事件发生于步阶处或步阶间的情况下才更新模型或即通过"步阶-事件"更新,基于事件的实现以有限的时间分辨率在基于步阶的模拟器中实现也是可能的。

[0095] 神经网络中的差分编码

[0096] 本公开的各方面涉及神经网络中的差分编码。

[0097] 在一些方面,神经网络学习或解决许多推断任务,包括对象分类、语音识别和手写识别。在许多应用中,神经网络从连续感官信息流中获取"感知"。作为示例而非限定,机器人(或智能电话)可以使用神经网络来提取图像序列上的高级特征或类别标签(即,图像分类)。在此类情境中,神经网络可以利用输入数据流的时间结构。由于该数据流在不同实例之间改变不大、或者以可预测的方式改变(例如,运动预测),因此本公开可以发送差分或差值结果而非在每个实例中发送所有数据值。本公开还可应用于针对机器学习网络的差分编码。例如,计算图像上的尺度不变特征变换(SIFT)特征可以使用基于与先前图像之差对SIFT值和位置的差分编码,或者可根据基于运动的前向估计。

[0098] 神经网络具有神经元层,其中底层表示原始数据并且较高层表示特征。底层可以是网络中的较低层,并且从底层接收输出的层可以是网络中的较高层。例如,"底"层可以是已经有一些预处理或初始特征提取的中间隐藏级,而"上"层可以是从该"底"层接收输入的层。当推断具有时间结构的感官流时,每个神经元可基于该神经元的激活历史来预测激活。在此类情形中,将激活值传播给其他神经元比发送实际激活值与基于历史的预测值之差(或误差)的效率更低。

[0099] 取决于预测有多好,神经网络的层级之间的通信得以减少。如果神经元之间的通信是二进制的(即,尖峰或无尖峰),则根据本公开的差值/误差办法通过差分编码来传播较

少的尖峰。由于预测值在神经元层接近100％的准确度，因此不太需要在较高层中的神经元处进行计算。如果神经元是非二进制的，则差分编码与发送激活值全集相比使用较少的位来达成相同精度水平。

[0100] 本公开可在神经网络中的层之间发送经编码值，其可以是预测激活值与激活值之差。此外，在本公开的一方面，可存在更改在神经网络中的层之间正被发送的信息的场合。激活值可以是差值，或者可以是激活值本身，或者可以是其他数据。对何种激活值、激活值差值、或一般值的确定可基于数种因素。这些因素包括激活值或激活值差值的位数、用来确定是否有任何数据在神经网络的层之间被发送的阈值、用来确定激活值的激活函数、接收到至输入神经元的输入、激活值的位宽、或其他因素。例如，可基于位数来设置阈值。即，在差值超过特定值时发送该差值，该特定值取决于可供特定神经元进行通信的位数。

[0101] 可以使用一个或多个激活函数来确定激活值和所预测的激活值。这些激活函数中的一者或多者可以是非线性函数。激活函数可以使用滤波器来实现，其还可确定激活值和/或经差分编码的激活值的编码。

[0102] 数据在神经网络内的发送或其他分发可在连续基础、周期性基础、或间歇性基础上进行。即，状态信息可以是跨网络周期性地(间歇性地)同步的。此外，激活值和/或经差分编码的激活值的编码可从接收到输入数据被延迟。

[0103] 尽管差分编码可以减少神经网络内传送的数据量，但是本公开还预见到设计选项可包括在网络内发送更多数据同时减少用于编码或确定要发送的数据的计算。例如，可以不发送关于激活值的预测，并且仅在接收到数据时通过系统转发实际数据。这种办法导致较大的数据吞吐量和较小的计算量。可在神经网络内的数据传输与数据计算之间进行设计折衷以满足各种神经网络设计。

[0104] 在神经网络内，用于一些神经元的激活函数中的一些激活函数可在神经网络的操作期间改变“模式”。此外，一些神经元可以总是在一种模式中操作，而其他神经元在另一种模式中操作。例如，一些神经元可仅发送经差分编码的数据，而其他神经元可发送整个激活值。一些神经元可在操作期间切换模式(例如，发送整个激活值直至某一时刻，并随后在该时刻之后发送经差分编码的激活值)。在神经网络内传送的数据的变化可基于神经网络内的数据分类、或其他因素，包括可用计算能力、传输可靠性、神经网络的大小、或其他约束。

[0105] 图5解说了根据本公开的某些方面的使用通用处理器502进行前述差分编码的示例实现500。与计算网络(神经网络)相关联的变量(神经信号)、突触权重、系统参数，延迟，频率槽信息，节点状态信息，偏置权重信息，连接权重信息，和/或激发率信息可被存储在存储器块504中，而在通用处理器502处执行的指令可从程序存储器506加载。在本公开的一方面，加载到通用处理器502中的指令可以包括用于以下操作的代码：在节点处接收输入事件，将偏置权重和连接权重应用于输入事件以获得中间值，基于中间值来确定节点状态，以及基于节点状态来计算表示后验概率的输出事件率以根据随机点过程来生成输出事件。

[0106] 图6解说了根据本公开的某些方面的前述差分编码的示例实现600，其中存储器602可以经由互连网络604与计算网络(神经网络)的个体(分布式)处理单元(神经处理器)606对接。与计算网络(神经网络)相关联的变量(神经信号)、突触权重、系统参数，延迟，频率槽信息，节点状态信息，偏置权重信息，连接权重信息，和/或激发率信息可被存储在存储器602中，并且可从存储器602经由互连网络604的(诸)连接被加载到每个处理单元(神经处

理器)606中。在本公开的一方面,处理单元606可被配置成在节点处接收输入事件,将偏置权重和连接权重应用于输入事件以获得中间值,至少部分地基于中间值来确定节点状态,以及基于节点状态来计算表示后验概率的输出事件率以根据随机点过程来生成输出事件。

[0107] 图7解说了前述差分编码的示例实现700。如图7中所解说的,一个存储器组702可与计算网络(神经网络)的一个处理单元704直接对接。每一个存储器组702可存储与对应的处理单元(神经处理器)704相关联的变量(神经信号)、突触权重、和/或系统参数,延迟,频率槽信息,节点状态信息,偏置权重信息,连接权重信息,和/或激发率信息。在本公开的一方面,处理单元704可被配置成在节点处接收输入事件,将偏置权重和连接权重应用于输入事件以获得中间值,至少部分地基于中间值来确定节点状态,以及基于节点状态来计算表示后验概率的输出事件率以根据随机点过程来生成输出事件。

[0108] 图8解说了根据本公开的某些方面的神经网络800的示例实现。如图8中所解说的,神经网络800可具有多个局部处理单元802,它们可执行本文所描述的方法的各种操作。每个局部处理单元802可包括存储该神经网络的参数的局部状态存储器804和局部参数存储器806。另外,局部处理单元802可具有用于存储局部模型程序的局部(神经元)模型程序(LMP)存储器808、用于存储局部学习程序的局部学习程序(LLP)存储器810、以及局部连接存储器812。此外,如图8中所解说的,每个局部处理单元802可与用于为该局部处理单元的各局部存储器提供配置的配置处理器单元814对接,并且与提供各局部处理单元802之间的路由的路由连接处理单元816对接。

[0109] 根据本公开的某些方面,每个局部处理单元802可被配置成基于神经网络的一个或多个期望功能性特征来确定神经网络的参数,以及随着所确定的参数被进一步适配、调谐和更新来使这一个或多个功能性特征朝着期望的功能性特征发展。

[0110] 在本公开的一方面,用于神经网络中的预测性差分编码的一般框架如下。人工神经元接收输入x(t)并发出输出y(t),其中t表示时间。输出y(t)可以是x(t)的非线性函数,诸如sigmoid(S型)函数:

[0111] $y(t)=\sigma(x(t))=e^{x}/(1+e^{x})$ (15a)

[0112] 或整流器非线性函数

[0113] $y(t)=\max(0,x(t))$。 (15b)

[0114] 神经元可以通过将S形表达式或其他表达式用作输出y(t)为1的概率来随机地发出二进制输出y(t)。

[0115] 输入x(t)可以是其他神经元的输出的加权线性组合:

[0116] $x_{i}(t)=\Sigma w_{ij}y_{j}(t)$, (16)

[0117] 其中$w_{ij}$表示第i个和第j个神经元的权重,并且j为连接至第i个神经元的所有神经元的索引。

[0118] 本公开的一方面允许预测性差分编码框架与任意人工神经元模型一起工作。向这些人工神经元添加状态变量允许神经元维持历史日志。函数s(t)标示一个或多个状态变量。每个神经元通过这些状态变量对历史进行跟踪,并且预测其将接收的输入$\hat{x}(t)$以及其将发出的输出$\hat{y}(t)$。这些预测减少了神经元之间的通信量(即,每个神经元现在仅发出预测误差$\delta y(t)=y(t)-\hat{y}(t)$而非实际输出y(t))。这些状态变量存储确定性模型的输入历史,并且在随机模型的情形中还存储输出历史。

[0119] 由于这些神经元发出预测误差，因此它们现在接收预测误差的加权组合z(t)而非如式(16)中的x(t)：

[0120] $z_i(t)=\Sigma w_{ij}\delta y_j(t)$。 (17)

[0121] 如果z(t)恰好等于输入x(t)的预测误差，则x(t)由下式准确地重构：

[0122] $x(t)=\hat{x}(t)+z(t)$。 (18)

[0123] 满足$\delta x(t)=x(t)-\hat{x}(t)$的条件为：

[0124] $z(t)=\delta x(t)$。 (19)

[0125] 在式16中，神经元接收到的是z(t)，而希望神经元接收的是δx(t)。如果满足式(19)，则该预测性差分编码方法与发出实际输出值的标准方法相比是准确但替换性的实现。即使式(19)未得到满足，该预测性差分编码方法也给出了近似实现。

[0126] 如果相同的线性函数被用于预测x(t)和y(t)，则式(19)得到满足。本公开的预测性差分编码方法变成替换性但准确的实现方法。如果预测是线性的，则近似应当是准确的。如果预测是非线性的，则近似可能不准确。

[0127] 神经元可以存储全部历史或者仅存储部分历史。令L表示神经元存储的历史量(即，神经元跟踪过去L个时间步阶上的输入和输出)。这使得状态变量函数：

[0128] $s(t)=\{x(t-1),y(t-1),x(t-2),y(t-2),\cdots,x(t-L),y(t-L)\}$。 (20)

[0129] 如果神经元输入-输出关系是确定性的(即，如果y(t)为x(t)的确定性函数)，则仅存储输入历史就足够了。对于任何给定输入，确定性算法计算具有唯一性输出值的数学函数，并且该算法产生该特定值作为输出。

[0130] 如果输入-输出关系是随机的，如受限波尔兹曼机(RBM)或深度置信网络(DBN)中那样，则输出历史也被存储。不同于确定性过程，随机过程具有一定的不确定性：即使初始条件(或起始点)是已知的，也会存在该过程可能演进的若干(通常无限多)方向。

[0131] 本公开还构想了在区域性和/或全局层面上使用差分编码框架。例如，对图像的运动向量估计可被用于确定整个图像或该图像的区域性部分的平均向量变化。全局或局部信息可被提供给所有神经元，并且随后突触前和突触后神经元两者可以使用这些区域性/全局差异来作出更好的预测。此外，突触后神经元可以提供差分反馈。在此方面，突触前神经元可将突触后神经元差分输出用于状态估计。这可减少神经网络中的层之间的通信量。

[0132] 给定历史s(t)，每个神经元使用以下线性滤波器来预测时间t处的输入和输出：

[0133] 对于k=1到L，$\hat{x}(t)=\Sigma\alpha_k x(t-k)$ (21)

[0134] 对于k=1到L，$\hat{y}(t)=\Sigma\alpha_{ky}(t-k)$。 (22)

[0135] 在此预测性框架内，式(19)得到满足。滤波器系数α1、α2、···、αL可随时间推移被学习或者被先验地选择。作为示例而非限定，如果L为1且α1=1(即，每个神经元将先前时间段输入和输出用作最佳预测)，则：

[0136] $\hat{x}(t)=x(t-1)$ (23a)

[0137] $\hat{y}(t)=y(t-1)$。 (23b)

[0138] 作为另一示例，如果L=2、α1=2、α2=-1(即，每个神经元假定输入正以线性方式改变)，则：

[0139] $\hat{x}(t)=2x(t-1)-x(t-2)=x(t-1)+(x(t-1)+x(t-2))$ (24a)

[0140] $$\hat{y}(t) = 2y(t-1) - y(t-2) = y(t-1) + (y(t-1) + y(t-2))。 \quad (24b)$$

[0141] 如果不同神经元使用不同的预测滤波器，则式(19)未得到满足。为了用本公开的差分编码方法得到准确预测，可能期望在整个神经网络上使用相同滤波器。

[0142] 每个神经元可分别将不同的x滤波器和y滤波器用于预测$\hat{x}(t)$和$\hat{y}(t)$。这些预测滤波器对于不同神经元可以相同或不同。为使神经元对“实际输入”的重构是准确的，x滤波器应当与所有扇入神经元的所有y滤波器匹配。确保匹配的一种方法是通过贯穿该网络使用相同的预测滤波器。在分层神经网络中，达成匹配的另一种方法是通过对一层中的所有神经元使用相同的x滤波器以及对先前层中的所有神经元使用相同的y滤波器。

[0143] 这些预测滤波器可以是固定的或是自适应的，并且可以是线性的或是非线性的。在关于非线性滤波器的情形中，关于每个输入突触的预测是期望的。对于线性滤波器，可以提供联合预测。在一个配置中，基线解决方案包括用于所有神经元的单个固定滤波器。另一种解决方案在线地估计滤波器系数值。在又一方面，滤波器可针对不同环境被配置或者甚至被优化，诸如用于室内相对于室外、或者静态相对于移动的滤波器集合。这些滤波器可通过以下方式来确定：具有预定义滤波器的码本、用于确定环境的方法、以及基于环境来选取特定滤波器。在另一方面，滤波器基于分类器(即，神经网络)的输出而被选择。

[0144] 在一个配置中，预定义滤波器是指数形状的。指数形状具有衰退因子，藉此迫使更多的增量更新。在一个示例中，提供了a.9系数。指数形状可减少或者甚至消除不稳定性和长期误差传播。指数形状还允许单步阶更新至将来值，从而仅在接收到非零输入时发生更新。应当注意，衰退因子与比特率之间存在折衷。即，较高的衰退因子将导致更多通信，而较低的衰退速率导致较少传输。在一个方面，不同神经元可将不同的衰退因子用于指数分布。在另一方面，滤波器是在线被学习的。例如，机器人在它正快速移动的情况下可以使用高衰退因子，并且在它正静止不动或者正缓慢移动的情况下可以使用低衰退因子。

[0145] 差分编码节省了用于神经元之间进行通信的资源。然而，差分编码也增加了开销。附加存储器存储状态变量或输入/输出值的历史。附加计算用于计算预测和预测误差。可通过使用指数滤波器形状来在一定程度上减少增加量。由此，差分编码的益处相对于附加开销之间存在折衷。

[0146] 差分编码可仅被用于神经元的子集。例如，考虑其中使用多个核(或机器)来仿真神经网络且不同核仿真不同神经元的场景。通信成本对于跨这些核或机器通信的那些神经元(即，这些神经元具有连接其他核中的神经元的输入突触或输出突触)较高。在此场景中，差分编码可仅被用于跨这些核或机器通信的神经元。而且，频繁改变的神经元可能不是用于差分编码的良好候选。不同神经元也可使用不同通信位宽或者甚至不同滤波器。在又一方面，神经元可以改变模式，其中在一种模式中它们发送差分更新而在另一种模式中它们发送实际结果。模式变化可基于触发，例如基于分类结果(即，来自神经网络的输出)是否令人满意。

[0147] 作为每个神经元预测输入和输出的替代，神经元的集合可以联合地预测其集体输入和输出。具体而言，在分层神经网络中，每层神经元可基于其向量输入和向量输出的联合历史来联合地预测向量输入和向量输出。线性预测性框架通过用矩阵替代标量滤波器系数来自然扩展到向量输入/输出场景(即，α1、α2、• • •、αL现在是矩阵)。

[0148] 联合预测具有胜过个体预测的优点的示例应用是视频上的视觉推断应用。考虑其

中人或物正在原本静态的环境中移动的场景。可以使用分层神经网络(诸如,深度卷积网络(DCN)),其中在一层中的神经元的集合表示空间响应图。在此情形中,滤波器矩阵基于从图像到图像的运动向量而被选择。这些运动向量可从标准运动估计技术自底向上获得,这些标准运动估计技术可从视频压缩文献中得到。或者运动向量可从DCN的输出自顶向下获得。注意,DCN可被训练成预测图像中的对象及其位置。神经元可基于附加全局输入(诸如运动向量)以及来自其正将信息发送至的神经元的反馈来进行预测。

[0149] 图9解说了根据本公开的各方面的用于在神经网络中执行差分编码的方法900。在框902,基于神经网络中的神经元的至少一个先前激活值来预测该神经元的激活值。在框904,基于神经网络中的该神经元的预测激活值与激活值之间的差值来对值进行编码。

[0150] 在一个配置中,用于差分编码的方法包括用于预测神经元的激活值的装置和用于对误差进行编码的装置。在一个方面,该预测装置和/或编码装置可以是被配置成执行所叙述的功能的通用处理器502、程序存储器506、存储器块504、存储器602、互连网络604、处理单元606、处理单元704、局部处理单元802、和/或路由连接处理单元816。在另一种配置中,前述装置可以是被配置成执行由前述装置所叙述的功能的任何模块或任何装置。

[0151] 本公开中所描述的神经网络可以是任何类型的神经网络,包括多层感知网络、深度卷积网络、深度置信网络、以及回流神经网络等。此外,尽管是关于神经元基于历史为其自身预测输入和输出且仅传播该神经元的输出中的误差来描述的,但是神经元可以使用其他神经元的误差和预测来预测它自己的输入和输出。

[0152] 以上所描述的方法的各种操作可由能够执行相应功能的任何合适的装置来执行。这些装置可包括各种硬件和/或软件组件和/或模块,包括但不限于电路、专用集成电路(ASIC)、或处理器。一般而言,在附图中有解说的操作的场合,那些操作可具有带相似编号的相应配对装置加功能组件。

[0153] 如本文所使用的,术语“确定”涵盖各种各样的动作。例如,“确定”可包括演算、计算、处理、推导、研究、查找(例如,在表、数据库或其他数据结构中查找)、探知及诸如此类。另外,“确定”可包括接收(例如接收信息)、访问(例如访问存储器中的数据)、及类似动作。而且,“确定”可包括解析、选择、选取、确立及类似动作。

[0154] 如本文所使用的,引述一列项目中的“至少一个”的短语是指这些项目的任何组合,包括单个成员。作为示例,“a、b或c中的至少一个”旨在涵盖:a、b、c、a-b、a-c、b-c和a-b-c。

[0155] 结合本公开所描述的各种解说性逻辑框、模块、以及电路可用设计成执行本文所描述功能的通用处理器、数字信号处理器(DSP)、专用集成电路(ASIC)、现场可编程门阵列信号(FPGA)或其他可编程逻辑器件(PLD)、分立的门或晶体管逻辑、分立的硬件组件或其任何组合来实现或执行。通用处理器可以是微处理器,但在替换方案中,处理器可以是任何市售的处理器、控制器、微控制器、或状态机。处理器还可以被实现为计算设备的组合(例如DSP与微处理器的组合、多个微处理器、与DSP核协作的一个或多个微处理器、或任何其他此类配置)。

[0156] 结合本公开所描述的方法或算法的步骤可直接在硬件中、在由处理器执行的软件模块中、或在这两者的组合中体现。软件模块可驻留在本领域所知的任何形式的存储介质中。可使用的存储介质的一些示例包括随机存取存储器(RAM)、只读存储器(ROM)、闪存、可

擦除可编程只读存储器(EPROM)、电可擦除可编程只读存储器(EEPROM)、寄存器、硬盘、可移动盘、CD-ROM,等等。软件模块可包括单条指令、或许多条指令,且可分布在若干不同的代码段上,分布在不同的程序间以及跨多个存储介质分布。存储介质可被耦合到处理器以使得该处理器能从/向该存储介质读写信息。在替换方案中,存储介质可以被整合到处理器。

[0157] 本文所公开的方法包括用于实现所描述的方法的一个或多个步骤或动作。这些方法步骤和/或动作可以彼此互换而不会脱离权利要求的范围。换言之,除非指定了步骤或动作的特定次序,否则具体步骤和/或动作的次序和/或使用可以改动而不会脱离权利要求的范围。

[0158] 本文中以及附录A和附录B中所描述的功能可以在硬件、软件、固件、或其任何组合中实现。如果以硬件实现,则示例硬件配置可包括设备中的处理系统。处理系统可以用总线架构来实现。取决于处理系统的具体应用和整体设计约束,总线可包括任何数目的互连总线和桥接器。总线可将包括处理器、机器可读介质、以及总线接口的各种电路链接在一起。总线接口可用于尤其将网络适配器等经由总线连接至处理系统。网络适配器可用于实现信号处理功能。对于某些方面,用户接口(例如,按键板、显示器、鼠标、操纵杆,等等)也可以被连接到总线。总线还可以链接各种其他电路,诸如定时源、外围设备、稳压器、功率管理电路以及类似电路,它们在本领域中是众所周知的,因此将不再进一步描述。

[0159] 处理器可负责管理总线和一般处理,包括执行存储在机器可读介质上的软件。处理器可用一个或多个通用和/或专用处理器来实现。示例包括微处理器、微控制器、DSP处理器、以及其他能执行软件的电路系统。软件应当被宽泛地解释成意指指令、数据、或其任何组合,无论是被称作软件、固件、中间件、微代码、硬件描述语言、或其他。作为示例,机器可读介质可包括随机存取存储器(RAM)、闪存、只读存储器(ROM)、可编程只读存储器(PROM)、可擦式可编程只读存储器(EPROM)、电可擦式可编程只读存储器(EEPROM)、寄存器、磁盘、光盘、硬驱动器、或者任何其他合适的存储介质、或其任何组合。机器可读介质可被实施在计算机程序产品中。该计算机程序产品可以包括包装材料。

[0160] 在硬件实现中,机器可读介质可以是处理系统中与处理器分开的一部分。然而,如本领域技术人员将容易领会的,机器可读介质或其任何部分可在处理系统外部。作为示例,机器可读介质可包括传输线、由数据调制的载波、和/或与设备分开的计算机产品,所有这些都可由处理器通过总线接口来访问。替换地或补充地,机器可读介质或其任何部分可被集成到处理器中,诸如高速缓存和/或通用寄存器文件可能就是这种情形。虽然所讨论的各种组件可被描述为具有特定位置,诸如局部组件,但它们也可按各种方式来配置,诸如某些组件被配置成分布式计算系统的一部分。

[0161] 处理系统可以被配置为通用处理系统,该通用处理系统具有一个或多个提供处理器功能性的微处理器、以及提供机器可读介质中的至少一部分的外部存储器,它们都通过外部总线架构与其他支持电路系统链接在一起。替换地,该处理系统可以包括一个或多个神经元形态处理器以用于实现本文所述的神经元模型和神经系统模型。作为另一替换方案,处理系统可以用带有集成在单块芯片中的处理器、总线接口、用户接口、支持电路系统、和至少一部分机器可读介质的专用集成电路(ASIC)来实现,或者用一个或多个现场可编程门阵列(FPGA)、可编程逻辑器件(PLD)、控制器、状态机、门控逻辑、分立硬件组件、或者任何其他合适的电路系统、或者能执行本公开通篇所描述的各种功能性的电路的任何组合来实

现。取决于具体应用和加诸于整体系统上的总设计约束，本领域技术人员将认识到如何最佳地实现关于处理系统所描述的功能性。

[0162] 机器可读介质可包括数个软件模块。这些软件模块包括当由处理器执行时使处理系统执行各种功能的指令。这些软件模块可包括传送模块和接收模块。每个软件模块可以驻留在单个存储设备中或者跨多个存储设备分布。作为示例，当触发事件发生时，可以从硬驱动器中将软件模块加载到RAM中。在软件模块执行期间，处理器可以将一些指令加载到高速缓存中以提高访问速度。随后可将一个或多个高速缓存行加载到通用寄存器文件中以供处理器执行。在参照以下述及软件模块的功能性时，将理解此类功能性是在处理器执行来自该软件模块的指令时由该处理器来实现的。

[0163] 如果以软件实现，则各功能可作为一条或多条指令或代码存储在计算机可读介质上或藉其进行传送。计算机可读介质包括计算机存储介质和通信介质两者，这些介质包括促成计算机程序从一地向另一地转移的任何介质。存储介质可以是能被计算机访问的任何可用介质。作为示例而非限定，此类计算机可读介质可包括RAM、ROM、EEPROM、CD-ROM或其他光盘存储、磁盘存储或其他磁存储设备、或能用于携带或存储指令或数据结构形式的期望程序代码且能被计算机访问的任何其他介质。另外，任何连接也被正当地称为计算机可读介质。例如，如果软件是使用同轴电缆、光纤电缆、双绞线、数字订户线(DSL)、或无线技术(诸如红外(IR)、无线电、以及微波)从web网站、服务器、或其他远程源传送而来，则该同轴电缆、光纤电缆、双绞线、DSL或无线技术(诸如红外、无线电、以及微波)就被包括在介质的定义之中。如本文中所使用的盘(disk)和碟(disc)包括压缩碟(CD)、激光碟、光碟、数字多用碟(DVD)、软盘、和蓝光®碟，其中盘(disk)常常磁性地再现数据，而碟(disc)用激光来光学地再现数据。因此，在一些方面，计算机可读介质可包括非瞬态计算机可读介质(例如，有形介质)。另外，对于其他方面，计算机可读介质可包括瞬态计算机可读介质(例如，信号)。上述的组合应当也被包括在计算机可读介质的范围内。

[0164] 因此，某些方面可包括用于执行本文中给出的操作的计算机程序产品。例如，此类计算机程序产品可包括其上存储(和/或编码)有指令的计算机可读介质，这些指令能由一个或多个处理器执行以执行本文中所描述的操作。对于某些方面，计算机程序产品可包括包装材料。

[0165] 此外，应当领会，用于执行本文中所描述的方法和技术的模块和/或其它恰适装置能由用户终端和/或基站在适用的场合下载和/或以其他方式获得。例如，此类设备能被耦合至服务器以促成用于执行本文中所描述的方法的装置的转移。替换地，本文所述的各种方法能经由存储装置(例如，RAM、ROM、诸如压缩碟(CD)或软盘等物理存储介质等)来提供，以使得一旦将该存储装置耦合至或提供给用户终端和/或基站，该设备就能获得各种方法。此外，可利用适于向设备提供本文所描述的方法和技术的任何其他合适的技术。

[0166] 将理解，权利要求并不被限定于以上所解说的精确配置和组件。可在以上所描述的方法和装置的布局、操作和细节上作出各种改动、更换和变形而不会脱离权利要求的范围。

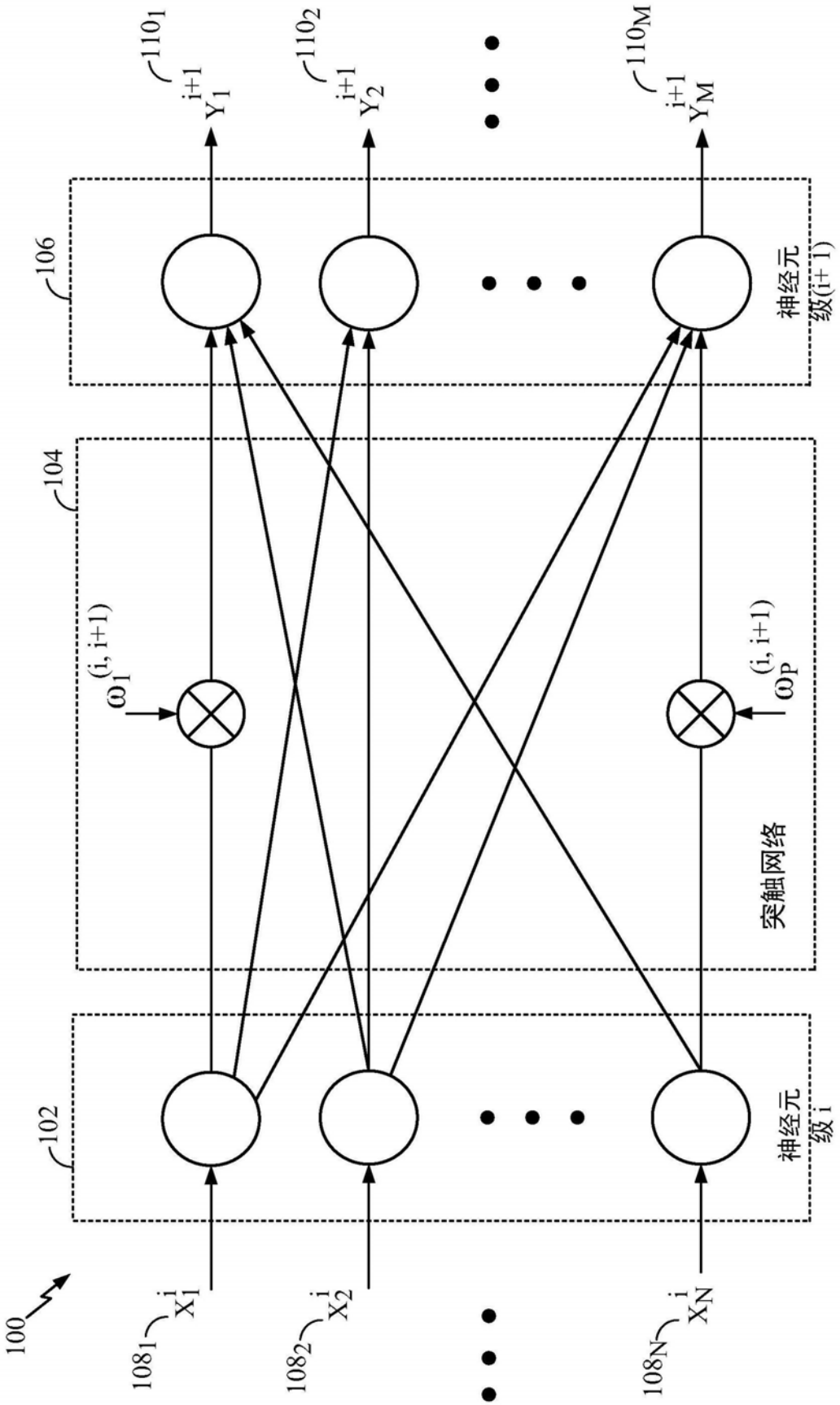

100
102
神经元级 i
104
突触网络
106
神经元级(i+1)
$108_1$
$108_2$
$108_N$
$X_1^i$
$X_2^i$
$X_N^i$
$110_1$
$110_2$
$110_M$
$Y_1^{i+1}$
$Y_2^{i+1}$
$Y_M^{i+1}$
$\omega_1^{(i,i+1)}$
$\omega_P^{(i,i+1)}$

图1

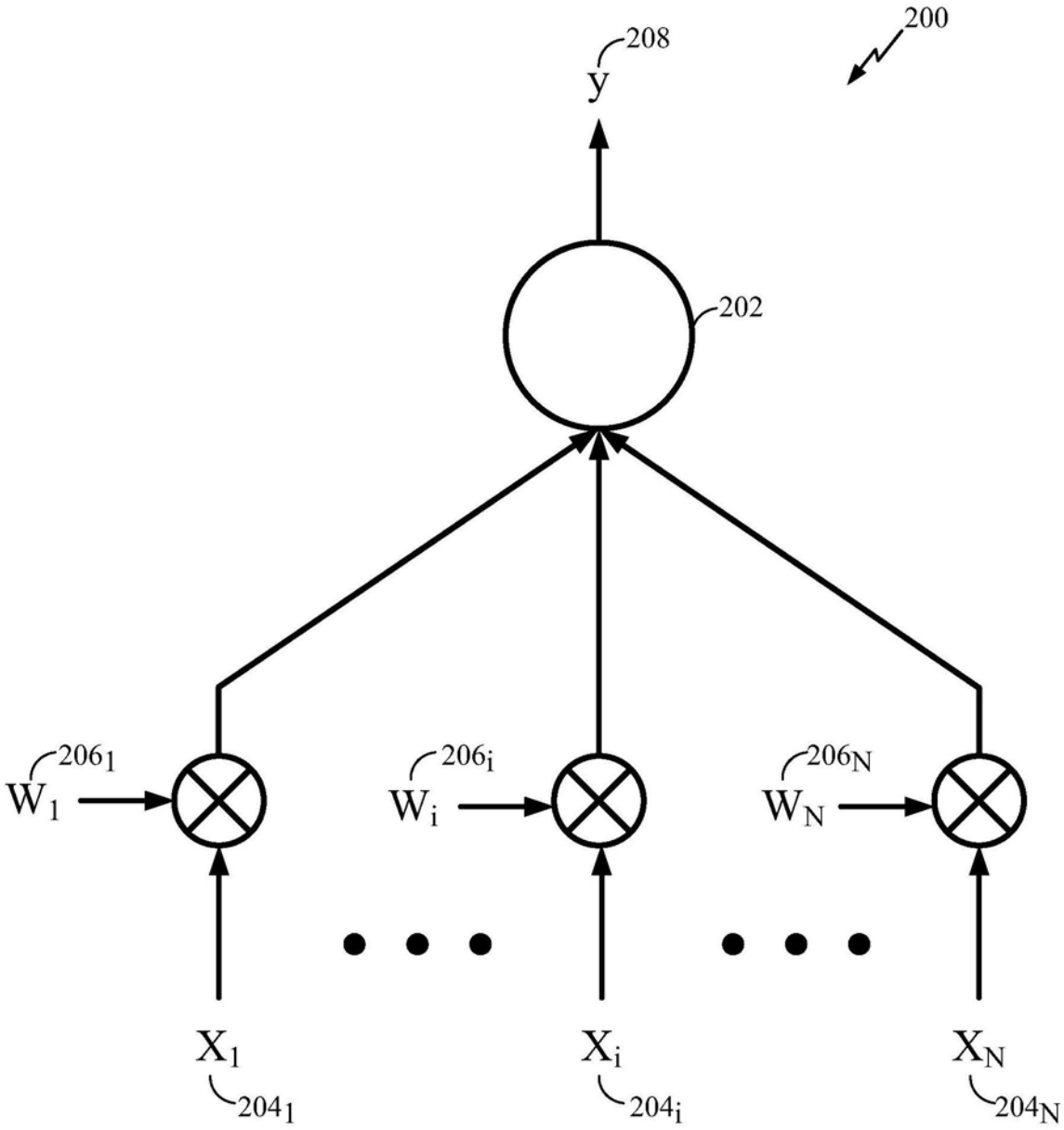



图2

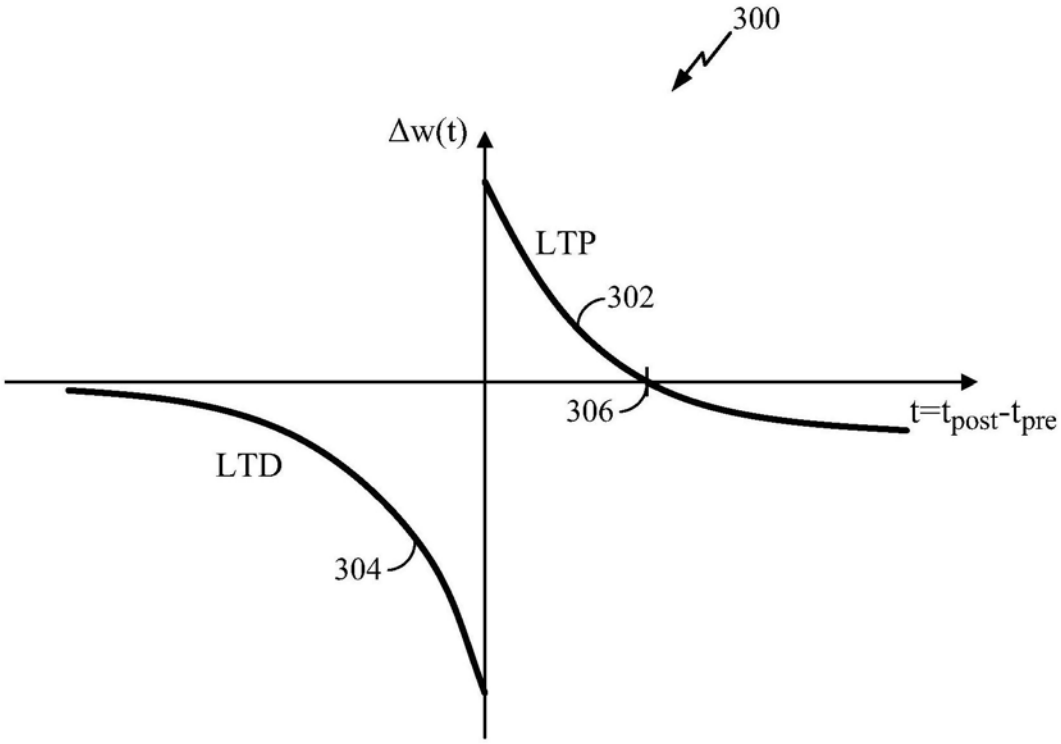

300
Δw(t)
LTP
302
306
t=t_post-t_pre
LTD
304

图3

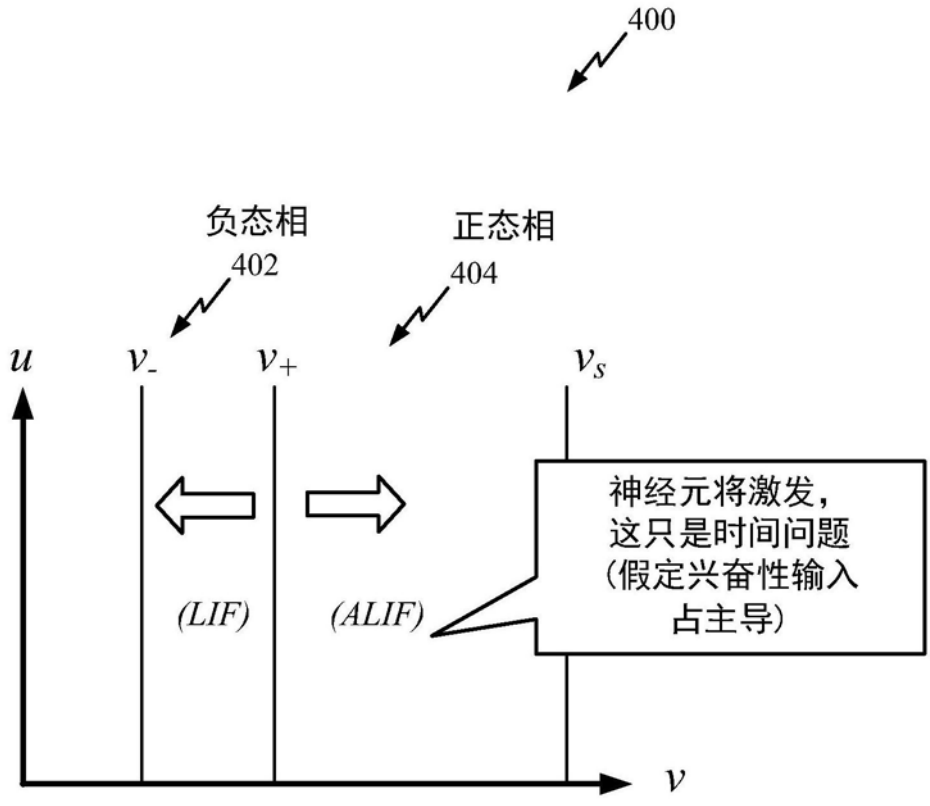

400
负态相
402
正态相
404
u
v-
v+
vs
神经元将激发，
这只是时间问题
(假定兴奋性输入
占主导)
(LIF)
(ALIF)
v

图4

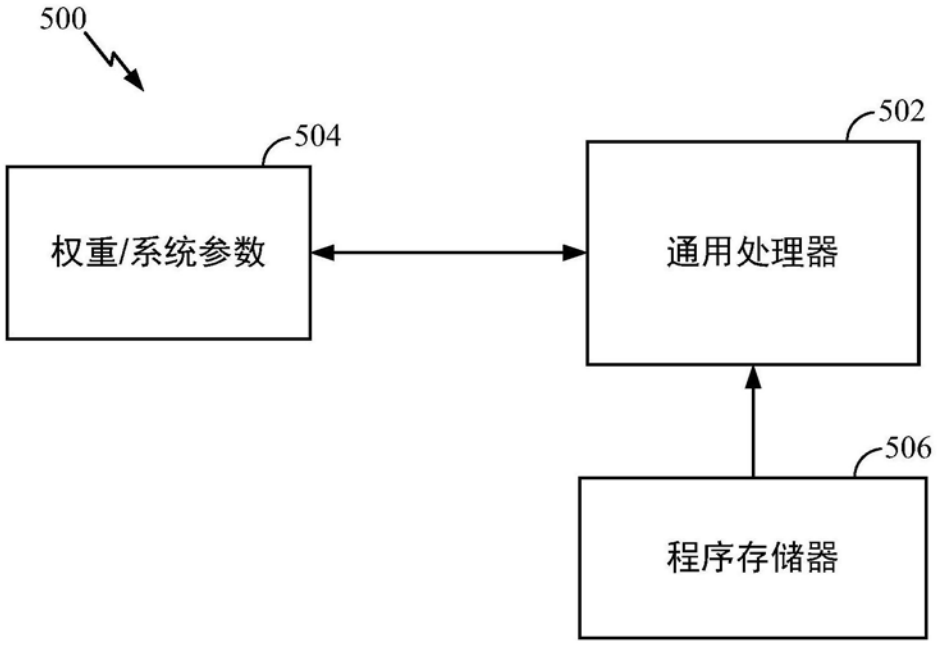

500
504
权重/系统参数
502
通用处理器
506
程序存储器

图5

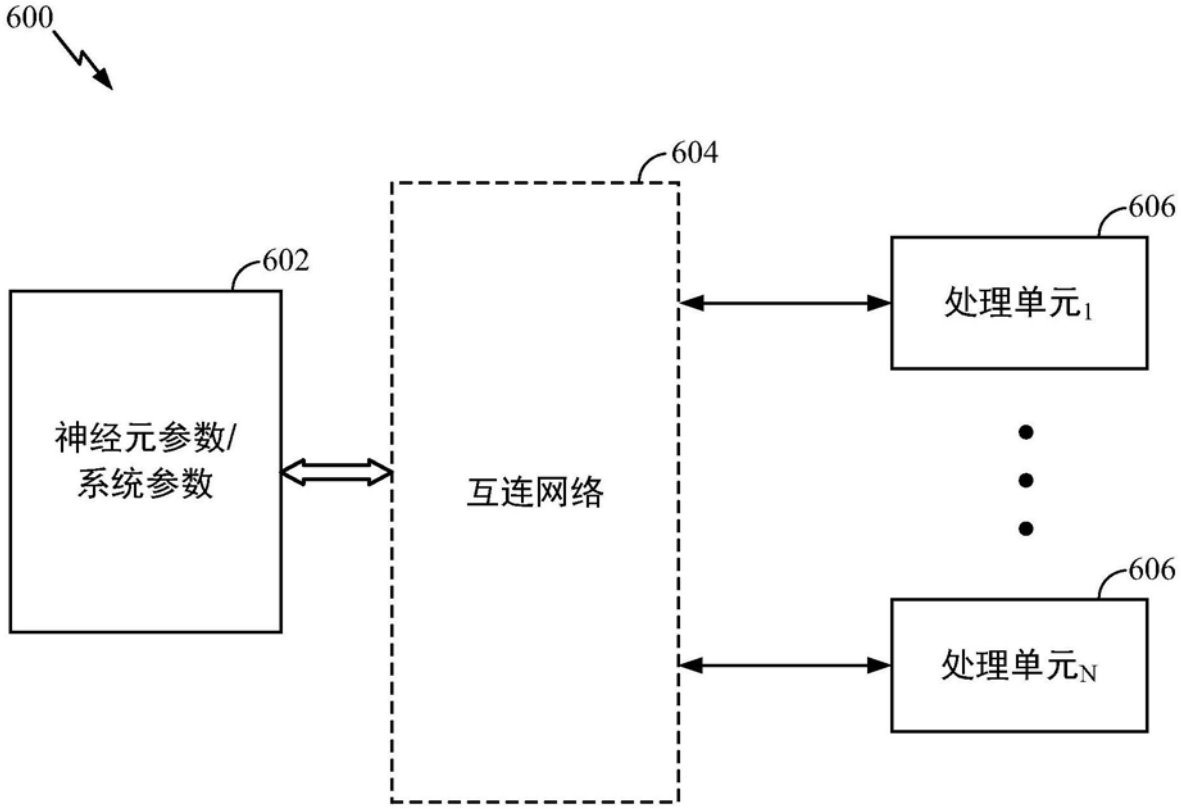

600
604
602
神经元参数/
系统参数
互连网络
606
处理单元1
606
处理单元N

图6

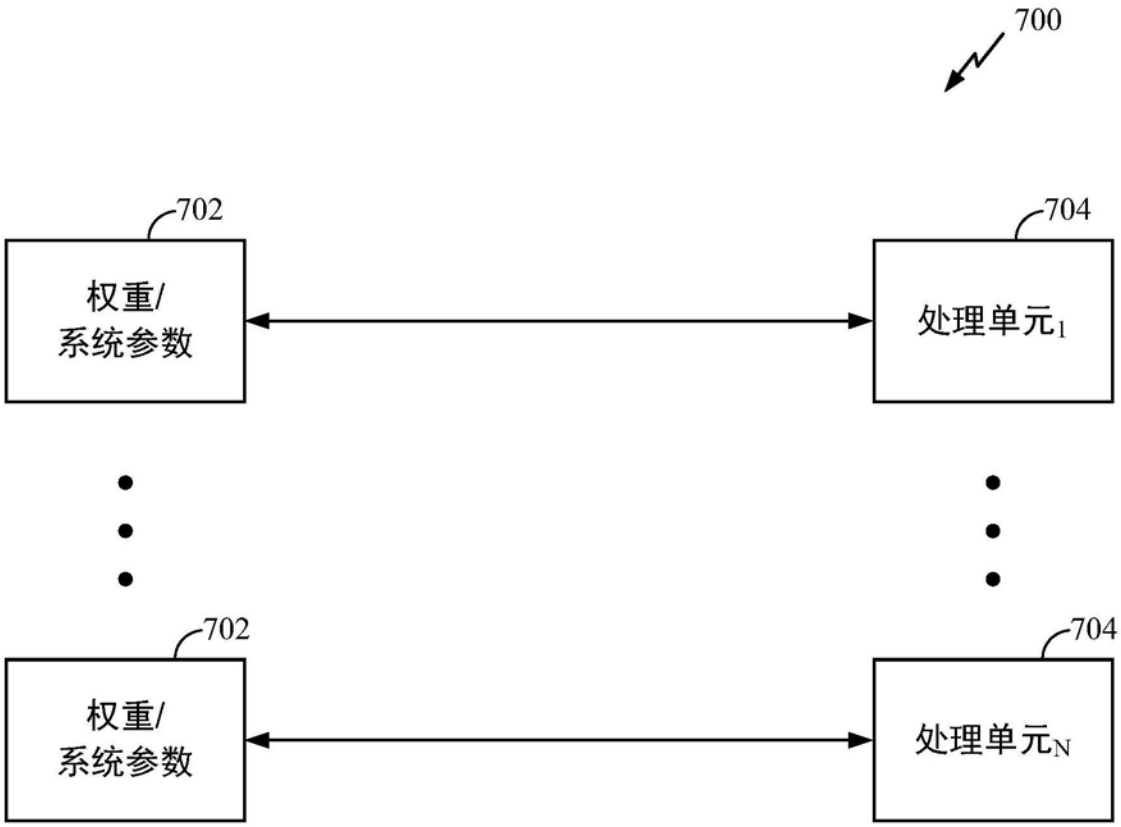



图7

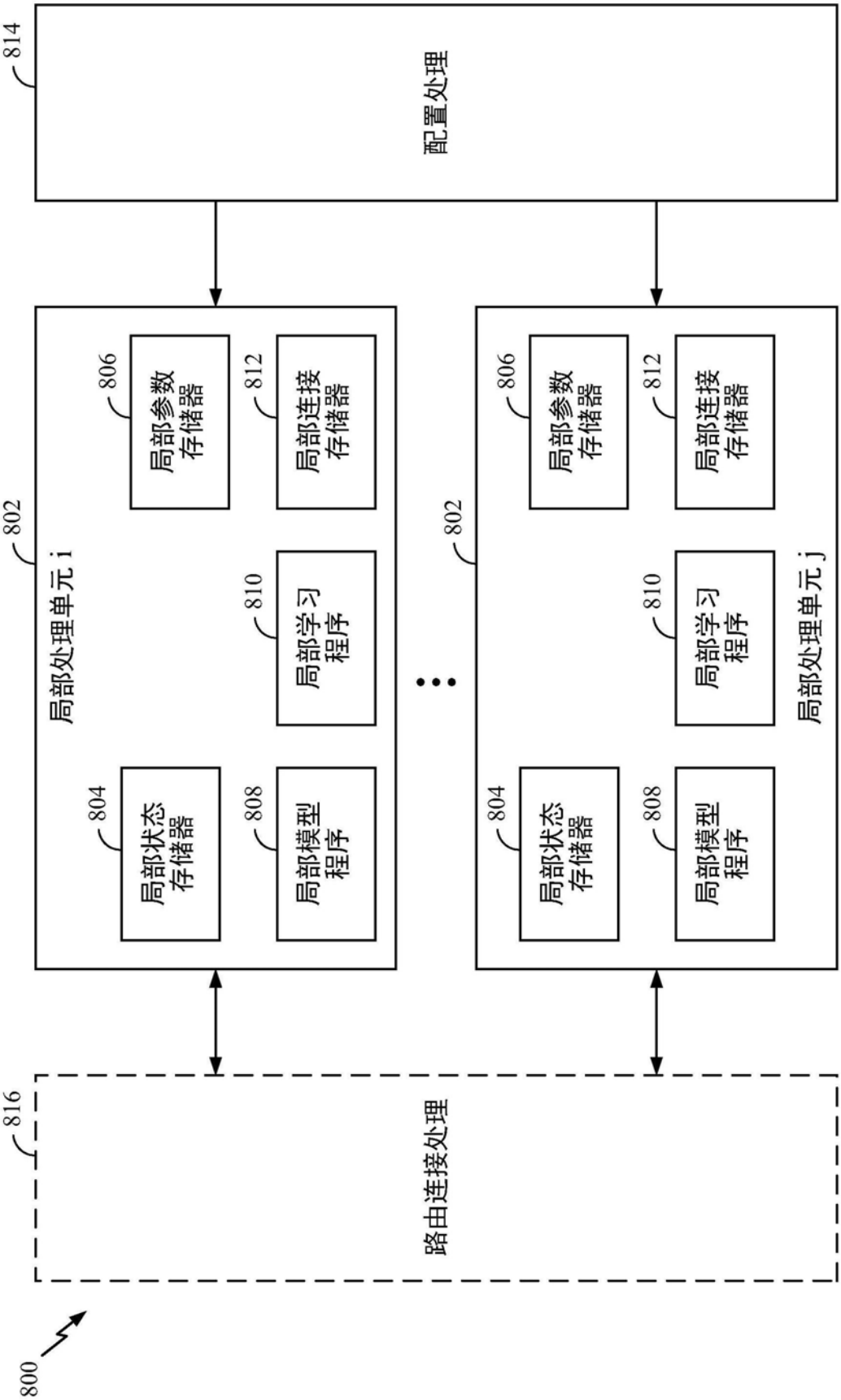

814
配置处理
802
局部处理单元 i
806
局部参数存储器
812
局部连接存储器
810
局部学习程序
804
局部状态存储器
808
局部模型程序
802
局部处理单元 j
806
局部参数存储器
812
局部连接存储器
810
局部学习程序
804
局部状态存储器
808
局部模型程序
816
路由连接处理
800

图8

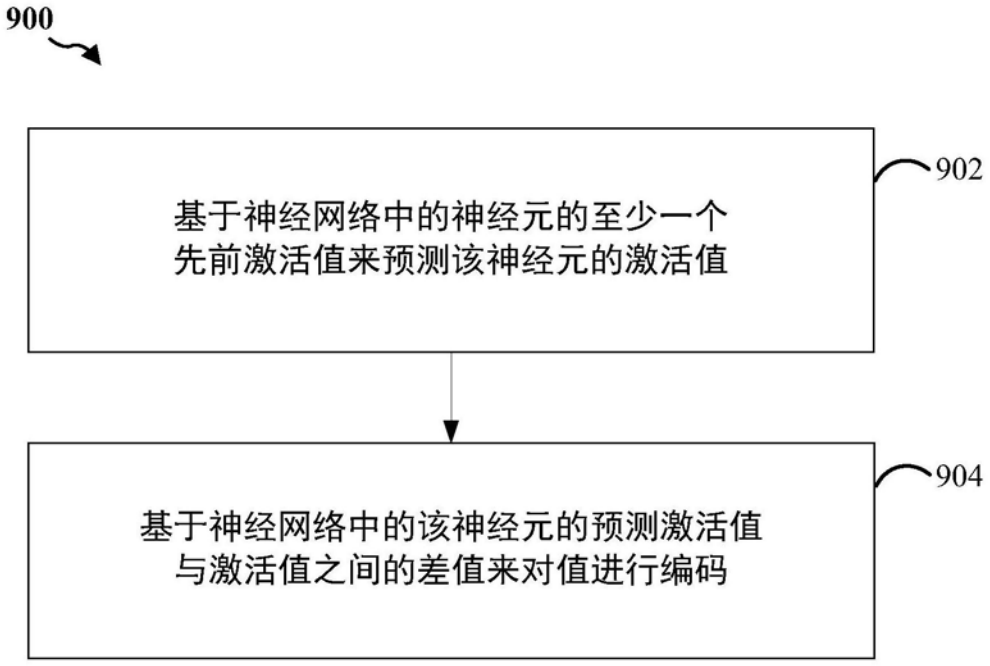

900
基于神经网络中的神经元的至少一个先前激活值来预测该神经元的激活值
902
基于神经网络中的该神经元的预测激活值与激活值之间的差值来对值进行编码
904

图9