计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2019.20190240
56(6):1135-1148,2019

类脑机的思想与体系结构综述

黄铁军[1] 余肇飞[1] 刘怡俊[2]
[1](北京大学计算机科学技术系 北京 100871)
[2](广东工业大学信息工程学院 广州 510006)
(tjhuang@pku.edu.cn)

Brain-like Machine: Thought and Architecture

Huang Tiejun[1], Yu Zhaofei[1], and Liu Yijun[2]
[1](*Department of Computer Science and Technology, Peking University, Beijing* 100871)
[2](*School of Information Engineering, Guangdong University of Technology, Guangzhou* 510006)

Abstract The theoretical limitation of the classical computing machinery, including all the computers with von Neumann architecture, was defined by Alan Turing in 1936. Owing to lack of the hardware neuromorphic devices, neural networks have been implemented with computers to realize artificial intelligence for decades. However, the von Neumann architecture doesn't match with the asynchronous parallel structure and communication mechanism of the neural networks, with consequences such as huge power consumption. To develop the neural network oriented architecture for artificial intelligence and common information processing is an important direction for architecture research. Brain-like machine is an intelligent machine which is constructed with neuromorphic devices according to the structure of biological neural network, and is better on spatio-temporal information processing than classic computer. The idea of brain-like machine had been proposed before the invention of computer. The research and development practice has been carried out for more than three decades. As one of the several brain-like systems being in operation, SpiNNaker focuses on the research on the architecture of brain-like systems with an effective brain-like scheme. In the next 20 years or so, it is expected that the detailed analysis of model animal brain and human brain will be completed step by step, and the neuromorphic devices and integrated processes will be gradually mature, and the brain-like machine with structure close to the brain and performance far beyond the brain is expected to be realized. As a kind of spiking neural networks, and with neuromorphic devices which behavior is true random, the brain-like machine can emerge abundant nonlinear dynamic behaviors. It had been proven that any Turing machine can be constructed with spiking neural network. Whether the brain-like machine can transcend the theoretical limitation of the Turing machine? This is a big open problem to break through.

Key words brain-like machine; spiking neural network (SNN); neuromorphic computing; Turing machine; synaptic plasticity

摘 要 经典计算机的理论边界在1936年就由图灵确定了,冯·诺依曼体系结构计算机也受限于图灵机模型.囿于神经形态器件的缺失,神经网络模型一直在经典计算机上运行.然而,冯·诺依曼体系结构与

收稿日期:2019-04-16;**修回日期**:2019-05-18
基金项目:国家自然科学基金项目(61425025);广东省重点领域研发计划项目(2018B030338001)
This work was supported by the National Natural Science Foundation of China (61425025) and the Key Research and Development Program of Guangdong Province (2018B030338001).

神经网络的异步并行结构及通信机制并不匹配,表现之一是功耗巨大,发展面向神经网络的体系结构,对于人工智能乃至一般意义上的信息处理都是重要方向.类脑机是仿照生物神经网络、采用神经形态器件构造的、以时空信息处理为特征的智能机器.类脑机的思想在计算机发明之前就提出了,研究开发实践也已经进行了 30 多年,多台类脑系统已经上线运行,其中 SpiNNaker 专注于类脑系统的体系结构研究,提出了一种行之有效的类脑方案.未来 20 年左右,预计模式动物大脑和人脑的精细解析将逐步完成,模拟生物神经元和神经突触信息处理功能的神经形态器件及集成工艺将逐步成熟,结构逼近大脑、性能远超大脑的类脑机有望实现.类脑机像生物大脑一样都是脉冲神经网络,神经形态器件具有真正的随机性,因此类脑机具备丰富的非线性动力学行为.已证明任何图灵机均可由脉冲神经网络构造出来,类脑机在理论上是否能够超越图灵机,是需要突破的一个重大问题.

关键词 类脑机;脉冲神经网络;神经形态计算;图灵机;突触可塑性

中图法分类号 TP389.1

1 类脑基本思想

1.1 图灵机

众所周知,现代计算机产生的数学基础是数理逻辑,物理基础是开关电路.数理逻辑的研究对象是证明和计算这 2 个直观概念符号化后的形式系统.1936 年阿兰·麦席森·图灵(Alan Mathison Turing,1912—1954)为了研究不可计算数而提出了图灵机模型,这一看似简单的思想实验抓住了数理逻辑和抽象符号处理的本质,划定了计算的理论边界:计算是机械式执行长度有限的算法的过程,这种计算都可以由图灵机完成;所有算法都可以编码成为一个整数,因此是可数的;尽管如此,并不存在枚举出所有算法的算法.

但是,现在很多人把计算这个概念随意泛化为任意的信息处理过程,这是不合适的.图灵机的状态和操作对象都是离散的,在图灵可计算意义下,1 和 0.1111…(无穷循环小数)是 2 个不同的数,产生这 2 个可计算数的图灵机也是不同的,图灵机并不能发现在极限意义下两者相等.极限是人类大脑的创造,在这个意义上,人脑是超越图灵机的.

1.2 冯·诺依曼体系结构

1938 年克劳德·艾尔伍德·香农提出开关电路模型,在数理逻辑和电路实现之间架起了桥梁.1946 年首台计算机 ENIAC 研制成功,实际上是一个近 1.8 万个电子管作为开关的大型开关电路系统.之前的 1945 年,参与了 ENIAC 项目的冯·诺依曼(John von Neumann,1903—1957)提出存储和计算分离的 EDVAC 结构,这篇报告分 15 章,长达百页,但是后来成为经典的"冯·诺依曼体系结构"只是其中的前 3 章,篇幅不到全文十分之一,之后报告重点就转到了神经系统.第 15 章没写完,冯·诺依曼后来也没继续写下去,而是转向研究怎样用不可靠元件设计可靠的自动机,以及建造自己能再生产的自动机.

冯·诺依曼体系结构被经典计算机沿用至今,虽然有各种优化,但无根本性变化.在摩尔定律作用下,冯·诺依曼体系结构计算机的性能呈指数增长,一直作为包括人工智能在内的各种信息应用的基础平台.2004 年至 2005 年前后,丹纳德尺度缩微定律(在半导体的尺寸不断缩小的同时其功耗密度大致保持不变)失效,普遍认为摩尔定律在持续 50 年后将于 2020 年左右走到尽头,迫使人们重新思考计算机的体系结构问题.

1.3 人工神经网络

1956 年,人工智能的概念正式登上历史舞台.60 多年来,人工智能经历了 3 次浪潮,基本思想可大致划分为三大流派:符号主义、连接主义和行为主义,从不同侧面抓住了智能的部分特征.连接主义也称神经网络学派,其基本思想是:既然人脑智能是由神经网络产生的,那就通过人工方式构造神经网络,进而产生智能.

神经网络思想的提出早于计算机的发明,1943 年麦卡洛克和皮茨把神经元想象成"全或无"的逻辑开关,他们提出的神经元模型至今还是人工神经网络使用的基本单元.80 多年来,人们提出了各种各样的人工神经网络,但是实现神经元和神经突触功能的物理器件一直未能发展起来.相比之下,冯·诺依曼体系结构计算机凭借集成电路摩尔定律的支持,性能呈指数增长,因此,缺少物理实现载体的人工神经网络逐步"寄生"在计算机上运行.但必须指出的是,人工神经网络结构和冯·诺依曼体系结构

毫无可比性,从体系结构角度看,冯·诺依曼体系结构不是实现神经网络运行的合理方案.

2006 以来,多层神经网络和机器学习相结合的深度学习在图像和语音识别等领域取得突破性进展,大规模深度神经网络和大数据训练对计算能力提出了更高需求,经典计算机运行神经网络能耗居高不下,按照神经网络的结构设计新的机器结构,已是大势所趋和必然选择.

1.4 生物神经网络

经典的人工神经网络(artificial neural network, ANN)借鉴了生物神经网络的基本特征,但过度简化.1)人工神经网络采用的神经元模型还是 1943 年提出的简化模型,与生物神经元的标准数学模型霍奇金-赫胥黎微分方程相距甚远,不是简单的数值计算;2)人类大脑是由数百种不同类型的上千亿的神经细胞所构成的极为复杂的生物组织,每个神经元通过数千甚至上万个神经突触和其他神经元相连接,即使采用简化的神经元模型,用目前最强大的计算机来模拟人脑,也还有 2 个数量级的差异;3)生物神经网络是一种复杂的脉冲神经网络(spiking neural network, SNN),采用动作电位表达和传递信息,按照非线性动力学机制处理信息,目前的深度学习等人工神经网络的时序特性还很初级.

仅就神经元模型而言,采用数字计算方法仿真生物神经元,计算复杂度比人工神经元模型要高多个数量级.即使采用简化的脉冲神经网络模型——泄漏积分发放(leaky integrate-and-fire, LIF)模型来实时仿真人类大脑,也约需要 100 台太湖之光超级计算机.

更严重的是,神经网络结构和冯·诺依曼体系结构大相径庭,这对性能的影响更为致命.2010 年左右提出的评价超级计算机的新指标——在大型随机图上每秒穿越的边数(traversed edges per second, TEPS),能够兼顾计算性能和通讯性能.如果将 2 个大脑神经元之间的一次脉冲传递类比为在图上穿越一个边,采用 TEPS 指标,人脑比当今最快的超级计算机也要快一个数量级.

与生物神经网络相比,人工神经网络过度简化,要实现更强的智能,需要更复杂、更精细的神经网络,最直接的蓝本就是生物神经网络.当前在计算机上采用软件方式仿真实现神经网络只是权宜之计,网络规模难以扩大,更直接的方案是直接按照神经网络结构设计全新的体系结构.

1.5 类脑机

理解意识现象和功能背后的发生机理(简称"理解智能")是人类的终极性问题,制造类似人脑的具有自我意识的智能机器(简称"制造智能")是工程技术领域重大挑战.一种常见看法是制造智能的前提是理解智能,这实际上把问题的解决建立在解决另一个更难问题的基础上,犯了本末倒置的错误.

要实现更强的机器智能乃至通用人工智能,首先要分清大脑的结构(主要是皮层神经网络)和大脑的功能(智能、意识)这 2 个层次.尽管目标是实现智能功能,但理解智能机器困难,更现实的做法是回到结构层次,尝试先制造出具有同样结构的机器,通过训练产生预期功能.自古以来人类的很多工程实践都是采用这种技术路线,以深度学习为例,其网络结构清晰、效果好,但机理不清楚,可解释性理论是下一步需要突破的问题,而不是设计深度神经网络的前提.

从人类大脑出发研究更强的机器智能乃至通用人工智能,我们认为更可行的技术路线是先结构仿脑,再功能类脑,最后才是理解大脑.因此,本文的"类脑",主要是指结构类脑,即仿真、模拟和借鉴大脑神经网络结构和基元(神经元、神经突触)信息处理过程,中心任务是制造类脑机(brain-like machine),或称神经机(neuromachine)[1-2].制造出这样的智能机器,理解机器智能的机理,将能加速对人类大脑智能奥秘的揭开.

类脑机是仿照生物神经网络、采用神经形态器件构造的、以时空信息处理为特征的智能机器.与生物神经系统一样,类脑机是一种脉冲神经网络,采用光电微纳器件模拟生物神经元和神经突触的信息处理功能,在仿真精度达到一定范围后,有望具备生物大脑类似的信息处理功能和系统行为.简言之,类脑机不是等待理解智能的机理后再进行模拟,而是绕过这个更为困难的科学问题,通过结构仿真等工程技术手段间接达到功能模拟的目的.

生物是类脑机的原型,生物智能活动主要是接收来自环境的多种刺激、实时处理并及时响应,这也是类脑机的主要功能和存在目的.

1.6 大脑解析进展

类脑机的体系结构源自生物大脑,这就需要获得生物大脑基本单元(各类神经元和神经突触等)的功能及其连接关系(网络结构).人脑拥有数百种、上千亿个神经元(即 10^{11} 数量级),每个神经元通过数千乃至上万神经突触和其他神经元相连接(连接

数量达到 10^{14} 数量级).尽管如此,人脑神经系统仍然是一个复杂度有限的物理结构,采用神经科学实验手段,从分子生物学和细胞生物学层次解析大脑神经元和突触的物理化学特性,理解神经元和突触的信号加工和信息处理特性,并无突破不了的技术障碍.

神经系统解析贯穿了神经科学百年历史.1906年,诺贝尔生理学或医学奖授予"在神经系统结构研究上的工作"的卡米洛·高尔基(Camilo Golgi,1843—1926)和圣地亚哥·拉蒙·卡哈尔(Santiago Ramon y Cajal,1852—1934),他们提出神经元染色法并绘制了大量精美的生物神经网络图谱,沿用至今.1939年剑桥大学阿兰·霍奇金和博士后安德鲁·赫胥黎开始研究神经元信号加工过程,自制工具测量到神经元的静息电位和动作电位.二战爆发,他们投笔从戎,1946年重新拿起膜片钳,精细测量神经元传递电信号(或称神经脉冲,更准确地称为动作电位)的动态过程,并给出了精确描述这一动力学过程的微分方程,称为霍奇金-赫胥黎方程(Hodgkin-Huxley方程,简称HH方程)[3].HH模型对不同类型的神经元具有通用性,1963年获得诺贝尔奖.

加拿大生理心理学家唐纳德·赫布1949年提出赫布法则(Hebb's Law):同时激发的神经元之间的突触连接会增强[4],至今这都是人工神经网络模型广泛采用的基本原则.1952年,中国现代神经科学奠基人张香桐(1907—2007)发现树突具有电兴奋性,树突上的突触可能对神经元的兴奋精细调节起重要作用,1992年国际神经网络学会授予张香桐终身成就奖,评价他"…为树突电流在神经整合中起重要作用这一概念提供了直接证据……为我们将来发展使用微分方程和连续时间变数的神经网络、而不再使用数字脉冲逻辑的电子计算机奠定了基础".1998年,Tsodyks和Markram等人提出了神经突触计算模型[5].同年,毕国强和蒲慕明提出了神经突触脉冲时间依赖的可塑性(spike-timing dependent plasticity, STDP)机制[6-7]:反复出现的突触前脉冲有助于紧随其后产生的突触后动作电位并将导致长期增强,相反的时间关系将导致长期抑制.2000年,宋森等人给出了STDP的数学模型[8-9].

2008年,美国工程院把"大脑反向工程"列为本世纪14个重大工程问题之一.2013年以来,欧洲"人类大脑计划"以及美、日、韩和我国的"脑计划"相继登场,都把大脑结构图谱绘制作为重要内容.2014年,"单细胞分辨的全脑显微光学切片断层成像"获得国家自然科学二等奖,并被欧洲人类大脑计划用作鼠脑仿真的基础数据.2016年3月,美国情报高级研究计划署(IARPA)启动大脑皮层网络机器智能(MICrONS)计划,对 1 mm^3 的大脑皮层进行反向工程,并运用这些发现改善机器学习和人工智能算法.2016年4月,全球脑计划研讨会(the Global Brain Workshop 2016)提出需要应对三大挑战,第一个挑战就是绘制大脑结构图谱[10]:"在10年内,我们希望能够完成包括但不限于以下动物大脑的解析:果蝇、斑马鱼、鼠、狨猴,并将开发出大型脑图谱绘制分析工具."2016年9月8日,日本东海大学宣布绘制出包括十多万神经元的果蝇大脑神经网络三维模型,2019年1月,《Science》封面文章报道只用了3天时间就对果蝇完整大脑进行了纳米级成像[11].

2018年,我国在北京怀柔开始建设"多模态跨尺度生物医学成像"国家重大科技基础设施,将具备从埃米到米、从微秒到小时跨越10个空间与时间尺度的解析能力,分步骤实现多种模式动物大脑的高精度动态解析.各方面的进展表明,人脑神经网络精细图谱有望在20年内完成.

2 类脑机研究进展

类脑机不是一个新想法.早在计算机发明之前的1943年,图灵和香农就曾围绕想象中的"电脑"进行过争论,香农提议把"文化的东西"灌输给电脑,而图灵高声反驳:"不,我对建造一颗强大的大脑不感兴趣,我想要的不过是一颗寻常的大脑,跟美国电报电话公司董事长的脑袋瓜差不多即可[12]."1950年,图灵在开辟人工智能方向的论文《计算机与智能》中明确表示:"真正的智能机器必须具有学习能力,制造这种机器的方法:先制造一个模拟童年大脑的机器,再教育训练[13]."冯·诺依曼也曾认真思考过大脑,根据他未完成的西列曼演讲整理而成的《计算机与人脑》一书1958年出版[14],上半部分为计算机,下半部分为人脑,讨论神经元、神经脉冲、神经网络以及人脑的信息处理机制.

实践意义上的类脑机研制可以追溯到20世纪80年代.美国生物学家杰拉尔德·艾德曼(Gerald Maurice Edelman,1929—2014)1981年提出了统称为"综合神经建模(synthetic neural modeling)"的理论,即逼近真实解剖和生理数据的神经系统大规模仿真[15],并研制了一系列名为"Darwin"的"仿脑机"(brain-based-devices, BBD)[16-17],通过从多种

仿真神经回路中进行选择而实现学习.起初是软件,1992 年开始采用硬件,以 2005—2007 年研制的达尔文 10 号和 11 号为例,仿真约 50 个脑区、10 万神经元和 140 万突触连接,通过模拟啮齿类动物走迷宫的过程,理解大脑空间记忆的形成过程.基于 BBD 的足球机器人于 2004—2006 年参加 RoboCup 机器人足球公开赛,曾 5 局全胜卡内基梅隆大学基于经典人工智能的系统.

现代微电子学和大规模集成电路先驱、加州理工学院教授卡弗·米德(Carver Andress Mead,1934—)也是在 20 世纪 80 年代把兴趣转向了生物神经系统的,与艾德曼关注神经元群体和神经环路不同,米德的关注点在神经元的硬件实现,开创了"神经形态工程(Neuromorphic Engineering)"这个方向[18-19],提出采用亚阈值模拟电路来仿真脉冲神经网络,并提出了"神经形态处理器(Neuromorphic Processors)"的概念.1989 年 5 月,米德在电路与系统研讨会(International Symposium on Circuits and Systems, ISCAS)会议期间组织了"模拟集成神经系统(Analog Integrated Neural Systems)"研讨会[20],主要参会人员至今仍然活跃在这一领域.

2.1 斯坦福大学的 Neurogrid 与 BrainStorm

米德 1989 年招收的博士生博阿汉(Kwabena Boahen)2005 年加入斯坦福大学,成立了"硅脑"(Brains in Silicon)实验室,2009 年研制出了神经形态电路板 Neurogrid,每块板 16 颗 Neurocore 芯片.每颗芯片内集成了 65 536 个神经元,每个神经元用 340 个亚阈值工作状态的晶体管模拟,这样一块 Neurogrid 板就支持 100 万个神经元和 60 亿个突触联结,能耗只有 5 W.每个 Neurocore 芯片都包括一个路由器,能够在其本地芯片、父芯片及其 2 个子芯片之间传送脉冲数据包.路由器支持多播树路由组织,其中脉冲数据被点对点传送到位于树中所有预期目的地之上的节点,然后到达所有目的地,需要时可以复制.据称 Neurogrid 在神经系统模拟方面可媲美能耗 1 MW 的超级计算机[21].

Neurogrid 团队 2017 年开发了新一代神经形态芯片 BrainStorm,这一项目 2013 年启动,由美国海军研究办公室资助,最后的成果将成为嵌入式应用和集群服务器上的计算芯片,可以运行全脑模型.目前还没有相关论文解释该项目的细节,但博阿汉指出 Brainstorm 与其他已有神经形态芯片设计存在着很大不同:"目前有很多神经形态设备使用的是超级计算机所使用的路由机制,就像网格一样.问题在于,在网格架构中你只能进行点对点信号传递.如果你想一次发出多个信号,系统就会锁死."博阿汉说 Brainstorm 是首个实现从高层次描述合成的脉冲神经网络的芯片,能解决多维非线性微分方程描述的问题,或者说是基于当前状态与输入随时间变化而变化的那类问题.

2.2 从软件仿真到 IBM TrueNorth 芯片

2005 年,瑞士洛桑联邦理工学院(EPFL)亨利·马克拉姆(Henry Markram,1962—)牵头"蓝色大脑计划",在 IBM 蓝色基因超级计算机上仿真大脑皮层[22].2007 年 IBM Almaden 研究中心认知计算研究组在美国国防高级研究计划局(DARPA)支持下开展神经形态自适应可塑性可扩展电子系统(systems of neuromorphic adaptive plastic scalable electronics, SyNAPSE)研究,开发了大脑模拟软件——皮层模拟器(cortical simulator),2009 年在蓝色基因超级计算机上实现了 8.61 T 个神经突触的猫脑模拟[23],所采用的神经元模型是简化的 LIF,即使如此,根据计算能力测算,实时模拟人类大脑也需要 100 台太湖之光超级计算机.同样在 2009 年,马克拉姆团队在蓝色基因超级计算机上构造出出生 2 周大鼠的新皮质柱精细模型,包括 1 万个神经元和数千万个突触连接,实现了生物神经网络才拥有的伽马振荡现象.在此基础上,由马克拉姆领衔的欧洲"人类大脑计划"于 2013 年 1 月获得欧盟批准,提出整合从单分子探测到大脑整体结构解析,实现全脑仿真模拟[24].

IBM 主导的 SyNAPSE 项目在超级计算机上进行大脑皮层仿真基础上,为了突破规模瓶颈,也开发了神经形态芯片 TrueNorth 芯片[25],2014 年 Science 将之列为年度十大科学进展.

TrueNorth 采用成熟的 CMOS 集成电路工艺,神经元采用简单的 LIF 模型,每片集成 4 096 个核,每核内有 256 个输入神经元和 256 个输出神经元,突触状态、神经元状态和参数、脉冲目的地址、轴突延迟等均用静态随机存储器记录.单片集成 100 万个神经元和 2.56 亿突触连接,耗费 54 亿个晶体管,单个芯片平均放电频率 20 Hz,单神经元放电功耗 26 pJ,芯片功耗低至 65 mW,大约是晶体管数量相当的传统 CPU 功耗的 $\frac{1}{5\,000}$.基于这款芯片,IBM 建立了 Corelet 编程模型、算法库和相应的软件开发环境,结合 Compass 模拟器,用户可以快速尝试不同的模型和参数,从中找出优化的方案.

2016 年 4 月,采用 TrueNorth,美国劳伦斯·利弗莫尔国家实验室和 IBM 公司公布了一款智能超级计算机,实验室数据科学副主任吉姆·布雷斯表示:“仿神经运算为我们创造了令人激动的新机会,这正是我们国家安全任务的核心——高性能运算和模拟技术的未来发展方向.仿神经计算机的潜在能力,以及它可以实现的机器智能,将改变我们研究科学的方式.”

2.3 欧洲的 SpiNNaker 和 BrainScaleS

为了实现全脑仿真的目标,欧洲人类大脑计划支持了 2 台大型神经形态计算系统的研制:英国曼彻斯特大学的 SpiNNaker 系统和德国海德堡大学的 BrainScaleS,2016 年 3 月 2 台阶段样机正式上线运行.

SpiNNaker[26-27]源于 2005 年开始的 EPSRC 项目,负责人是 ARM 处理器发明人史蒂夫·佛伯(Steve Furber,1953—).SpiNNaker 系统采用定制 ARM 处理器作为基本单元,分为 5 代,最初的 102 机使用了约 10^2 个 ARM 核,计划 2020 年完成的 106 机则集成了约 10^6 个 ARM 核.SpiNNaker 研究的中心任务就是探索新的体系结构,采用包交换来模拟神经元之间的异步稀疏脉冲交换,可以在物理连接大大少于大脑的情况下实现相同性能的信息交换,具体细节将在第 3 节详细介绍.

BrainScaleS 由德国海德堡大学卡尔海因茨·迈耶(Karlheinz Meier,1955—2018)教授负责[28-29],前身是 2005—2010 年的 FACTES 项目,特点是从微观层面研究神经元的信号处理特性及模拟电路实现,在介观层面研究突触可塑性及数字电路实现,在 8 英寸晶园上实现了 20 万神经元和 5 千万突触,晶圆内总线速度达每秒 1 T 脉冲,晶圆间分布式通信速度每秒 10 G 脉冲.在人类大脑计划支持下,2016 年完成了 20 块晶圆、400 万神经元和 10 亿突触的神经形态计算系统[30],速度比生物系统快 1 万倍.2022 年(也就是人类大脑计划结束前)预计构造出一个 500 块到 5 000 块晶圆组成的大型系统,即使是 500 块方案,也能同时仿真 5 亿神经元,由于其速度比生物神经元高万倍,因此将具备实时仿真人类大脑的能力.

2.4 我国相关进展

我国类脑研究起步较晚,但近年来十分活跃,北京大学、清华大学、中国科学院自动化研究所、浙江大学、四川大学等单位成立了多个类脑计算或类脑智能方面的研究中心.

2015 年 9 月 1 日,北京市科学技术委员会正式发布“北京脑科学研究”专项规划,从“脑认知与脑医学”和“脑认知与类脑计算”2 个方面进行布局[31].“脑认知与类脑计算”沿着“结构仿真、器件逼近和功能超越”这条技术路线,布局了 3 个层次、9 个方面的科研任务:建设四大基础性公共平台(大脑解析仿真平台、认知功能模拟平台、神经形态器件平台和类脑计算机系统平台),开发 2 款类脑计算处理器芯片(类脑处理器和机器学习处理器),研制类脑计算机软硬件系统,在视听感知、自主学习、自然会话三大类脑智能方向取得突破并实现规模应用.经过 3 年多的持续支持,北京已经在类脑计算方面形成了较为系统的技术积累,清华大学研制的天机系列芯片和北京大学研制的超速全时视网膜芯片是其中的代表性成果.

清华大学团队提出了类脑混合计算范式架构,开发了“天机”系列类脑芯片.2015 年 11 月研制出首款跨模态异构融合神经形态类脑计算芯片,可进行大规模神经元网络的模拟,具有超高速、实时、低功耗等特点,相关结果于 2016 年 12 月发表在《Science》智能机器人特刊.2017 年 10 月研制成功天机 2 代神经形态芯片,采用 28 纳米半导体技术,集成了千万突触和约 4 万个神经元,同时支持脉冲神经网络算法和人工神经网络算法,与 IBM TrueNorth 相比,在芯片密度、速度和带宽都有大幅度提升.2018 年利用脉冲神经网络的时空特性,实现了在时空域的 SNN 误差反向传播算法,解决了函数逼近的方法处理脉冲发放时刻不可导问题,建立了 SNN 全连接及卷积神经网络新算法.

北京大学在北京“脑认知与类脑计算”支持下,围绕视觉系统解析仿真开展研究,研制出类脑机的“眼睛”.2015—2016 年对灵长类视网膜进行了高精度解析仿真,实现了视网膜中央凹神经细胞和神经环路精细建模,提出了模拟视网膜机理的仿生视频脉冲编码模型.2017—2018 年初,研制成功脉冲阵列式超速全时仿视网膜芯片.生物视觉信息处理机制虽然优越,但受限于生理限制,“主频”很慢,灵长类视网膜每秒发放的神经脉冲数平均不超过数十个.仿视网膜芯片脉冲发放频率达到 40 000 Hz,“超速”人眼千倍,能够“看清”高速旋转叶片的文字.“全时”是指从芯片采集的神经脉冲序列中重构出任意时刻的画面,这是真正机器视觉的基础,有望重塑包括表示、编码、检测、跟踪、识别在内的整个视觉信息处理体系.

浙江大学及杭州电子科技大学联合研究团队主要面向低功耗嵌入式应用领域,于2015年研发了一款基于CMOS数字逻辑的脉冲神经网络芯片"达尔文",支持基于LIF神经元模型的脉冲神经网络建模.2016 IEEE CIS计算智能相关的暑期学校将达尔文芯片作为一个案例供所有参加人员编程实践与应用开发.

2017年国家自然基金委信息科学部研究确定了"人工智能(F06)"代码,专门设置了"认知与神经科学启发的人工智能(F0607)"方向,其中与类脑直接相关的支持方向包括:视听觉感知模型、神经信息编码与解码、神经系统建模与分析、神经形态工程、类脑芯片、类脑计算.从2018年起,我国类脑领域的基础研究已经全面展开.

3 脉冲神经网络体系结构SpiNNaker

SpiNNaker是脉冲神经网络体系结构(The spiking neural network architecture)的缩写,是英国曼彻斯特大学Steve Furber教授带领的先进处理器技术团队(APT)研发的类脑计算系统,研究始于2005年,目的是借鉴大脑神经网络结构研究新的计算体系结构.

SpiNNaker是一个大型脉冲神经网络,采用独特的全局异步局部同步(GALS)互连网络结构,最新系统将不同时域的一百万ARM微处理器核心和1 200个互连计算主板高效集成为1台高度并行的超级计算机,每秒执行200万亿次定点运算操作,支持实时事件驱动的编程模式,适用于生物神经网络的实时模拟.

3.1 体系结构

SpiNNaker系统由ARM微处理器核心、多核CPU芯片、计算主板、机架、机柜和整机等6个不同的层次构成,如图1所示.2018年11月上线的最新系统采用的微处理器核心是200 MHz的32-bit ARM968微处理器,拥有32 KB指令存储器和64 KB数据存储器,不带浮点运算单元.每颗多核CPU芯片包含18颗ARM968和一个中央片上网络路由器,用异步片上网络连接.48颗SpiNNaker多核CPU芯片构成一块计算主板.24块主板组成一个机架.5个机架构成一个机柜.10个机柜组成整个SpiNNaker系统.因此,SpiNNaker系统包含的ARM CPU数量为18×48×24×5×10=1 036 800个.一个200 MHz的ARM CPU可以生物实时模拟1 000~10 000个IF(integrate-and-fire)级别的简单神经元模型,整个SpiNNaker系统理论上可以生物实时模拟10~100亿个这样的神经元.

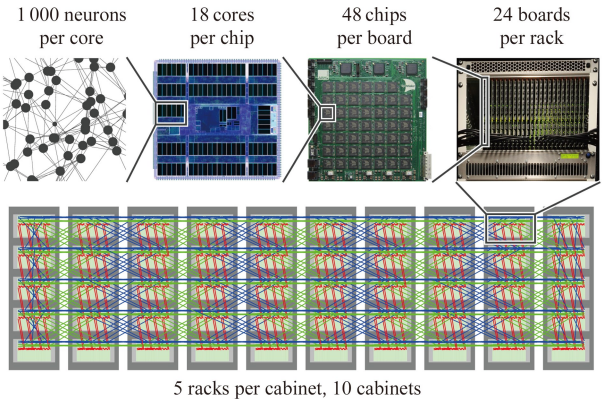
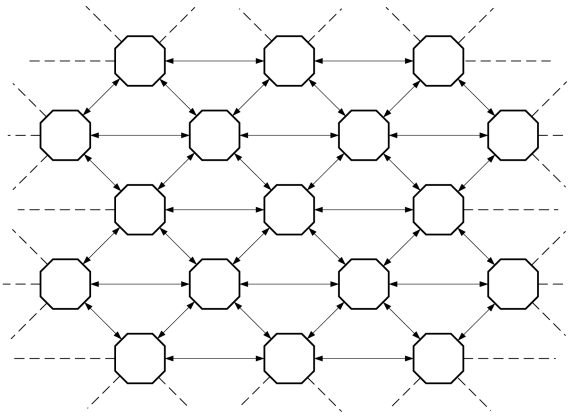


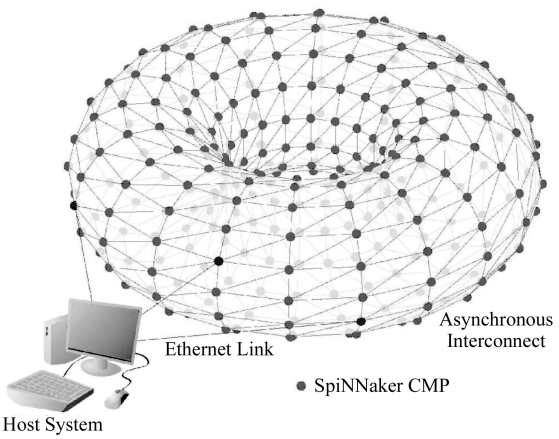
Fig. 1 Hierarchical structure of SpiNNaker

图1 SpiNNaker系统的层次结构

SpiNNaker多核CPU芯片包含一个6端口的通信路由器与其他芯片连接.整个系统形成一个六边形平面网格(hexagon 2D mesh)的拓扑结构,如图2(a)所示,平面网格结构的左右和上下方向的边缘接口相连,最终构成一个轮胎形态的环状结构,如



(a) Hexagonal mesh topology



(b) Hexagonal mesh topology of SpiNNaker

Fig. 2 Network topology of the SpiNNaker system[32]

图2 SpiNNaker系统的网络拓扑结构[32]

图 2(b)所示(其中 CMP 为 chip multiple processors 的缩写,表示多核处理器芯片).

3.2 海量脉冲异步传输机制

人脑中的每个神经元通过神经突触与成千上万其他神经元相连接,每个神经脉冲要传递给成千上万个神经元,这种高扇出(fan-out)的多播(multicast)传输方式对传统超级计算机来说是个巨大挑战.传统超算支持点对点的大数据块传输非常高效,但实现海量短小神经脉冲数据包的多播传输效率很低.

SpiNNaker 系统研发了一种适用于大规模脉冲神经网络模拟的高效"源地址多播传输"机制.SpiNNaker 支持相邻神经元数据包(nearest neighbor package)传输、点对点数据包(point-to-point package)传输、固定路径(fixed route package)传输、神经脉冲数据包(the neural event package)传输等 4 种不同的数据包传输方式.前 3 种数据包用于初始化、状态检测、控制信息和参数传递等.

神经脉冲数据包传输是 SpiNNaker 中最重要的传输方式,其数据包的格式如图 3 所示.神经脉冲数据包为 40 b 或 72 b,包括 32 b 数据载荷和 8 b 控制字,还可以额外携带一个 32 b 的数据载荷.通常意义上的神经脉冲数据包不需要携带数据,一个数据包的到来代表着一个神经脉冲到来的事件.在 SpiNNaker 系统中,负载是发送该神经脉冲的神经元的 32 b 源地址(可能遵循一定的规则,如 16 b 代表 CPU 编码,16 b 代表在该 CPU 中模拟的神经元地址).然而,这个结构中并没有说明目的地址,数据包如何被精确地传递到成千上万个目的地呢?SpiNNaker 提出了"源地址多播传输"机制——数据包经过的路由器根据 32 b 源地址查找路由表,将该数据包复制到不同的输出口,一级级传递下去.

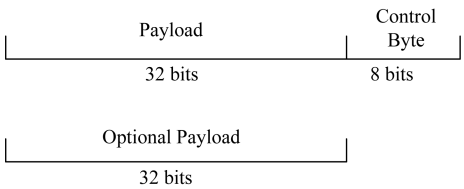
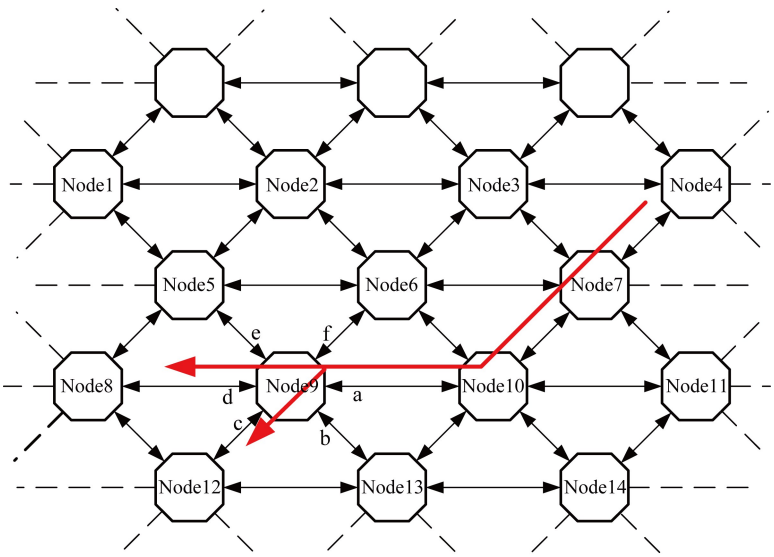


Fig. 3 Neural spike package layout[32]

图 3 神经脉冲数据包结构[32]

如图 4 所示,每颗多核 CPU 芯片的路由器有东



(a) An example of source-addressed multicasting routing

Source Address	Output Port
5	e
2	a
1	b
3	f
8	f
⋮	⋮

Routing Table of Node7

Source Address	Output Port
5	e
4	d, c
8	f, a, d
11	e
7	local
⋮	⋮

Routing Table of Node9

Source Address	Output Port
4	d
2	a
1	b
3	a
8	f
⋮	⋮

Routing Table of Node10

Note: Shaded values represent effective routing information.

(b) Examples of routing maps

Fig. 4 Propagation mechanism of source-addressed multicasting

图 4 源地址广播的实现机制

(E)、西(W)、东南(SE)、西南(SW)、东北(NE)和西北(NW)6个传输方向,简称a,b,c,d,e,f,还有一个“local”方向表示该芯片本地18颗CPU的传输方向.路由器中有一个路由表.路由表在实现上是一个CAM芯片,存储许多行(如1 024行)的路由信息,支持所有行源地址的并行比较.每行有左右2列,左边代表一个发送数据包的源地址,右边代表传输的方向,由多位构成,表示多个方向.

路由器接收到一个数据包后,根据包的源地址并行查找路由表.如果查到,路由器就按照路由表的方向指示向一个或多个方向传输该数据包;如果没有查到,路由器将包按照来的方向直接传递(a→d,b→e,c→f,d→a,e→b,f→c).如图4所示,假如节点4传递一个数据包到节点7,源地址为4;节点7路由表中没有4,按照直线传给10;节点10路由表中显示源地址为4的数据包按照d方向传递给9;节点9路由表再按照d,c方向传递给节点8和节点12(这时数据包被复制了),以此类推.路由表的信息在脉冲神经网络运行之前由sPyNNaker软件系统(3.3节介绍)统一初始化.

采用源地址多播传输机制进行海量神经脉冲的高扇出分发传递,使得数据包无需指明众多的目的地址就能够大规模地按照指定方向进行多播并行传输,有利于保证数据包格式的规范性,大大缩短了数据包的长度,提高了传输的速度.

3.3 SpiNNaker软件系统

SpiNNaker的软件系统称为sPyNNaker,可以将PyNN语言描述的脉冲神经网络解析并在SpiNNaker系统中仿真运行.PyNN语言是一种基于Python的跨平台脉冲神经网络描述高级语言,支持主流的脉冲神经网络软件仿真平台,包括NEST,NEURON和Brian,因此SpiNNaker和BrainScaleS都支持它,以兼任支持各种脉冲神经网络模型.

sPyNNaker软件系统架构如图5所示,各部分组成和功能均有显示.

1) Front End Interface:PyNN.PyNN的前端接口模块,用户可以在客户端利用PyNN接口编写脉冲神经网络模型.

2) Mapping:Placement, Partitioning, Routing, Data Gnenration.将PyNN描述的脉冲神经网络根据用户分配的硬件资源分解并映射到相应的CPU、内存和路由表中,生成配置信息.

3) Python Interface to SpiNNaker Hardware.负责客户端与SpiNNaker硬件系统的接口,包括将配置信息通过互联网下载到SpiNNaker计算机、传输模拟控制命令、将SpiNNaker的模拟结果传回到前端等功能.

4) Visualization. SpiNNaker的虚拟可视化界面.

5) SARK(SpiNNaker application runtime kernel).底层的硬件管理,主要控制DMA、网络接口和通信控制器等.

6) Event-Driven SpiN1API.支持事件驱动的操作系统,主要负责维护CPU内核中的任务安排进程、

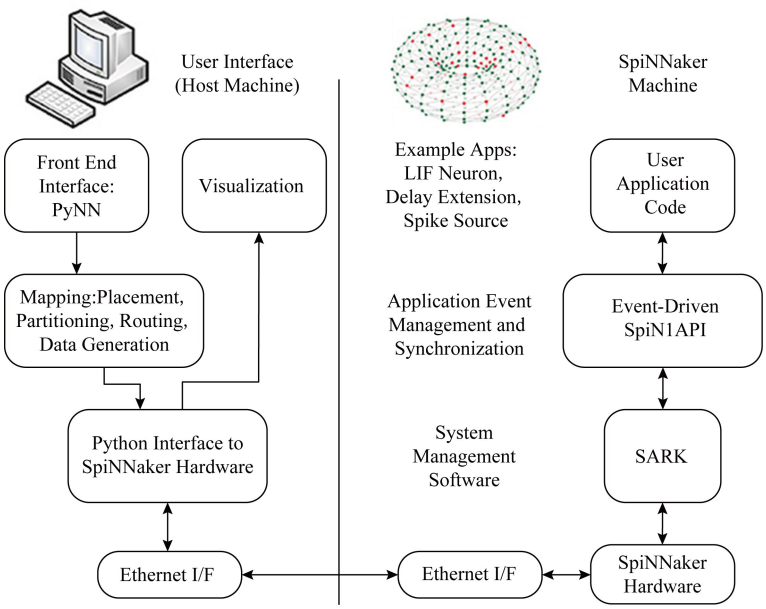


Fig. 5 Software system architecture of SpiNNaker[32]

图5 SpiNNaker的软件系统结构[32]

任务调度进程和快速事件响应等 3 个主要进程,支持实时事件模拟.

7) User Application Code.与神经网络生成和模拟相关的应用开发库文件,支持不同神经元模型、延时模型和脉冲源等.

sPyNNaker 采用时间驱动(time-driven)和事件驱动(event-driven)两种混合驱动方式来模拟脉冲神经网络,时间驱动模拟神经元的变化,事件驱动模拟突触的变化,如图 6 所示.CPU 用时间轮询的方式模拟生物时间,例如 ΔT CPU 时间模拟 1 ms 的生物时间,在 ΔT 时间内,CPU 轮询该 CPU 中模拟的所有神经元,更新它们的状态.CPU 同时还需要响应来自于本 CPU 中神经元或者外部神经元发送的神经脉冲到达的事件,支持以组播的方式更新相连接的突触,并更新神经元的状态.各神经元状态变化后可能产生新的神经脉冲,触发事件驱动.

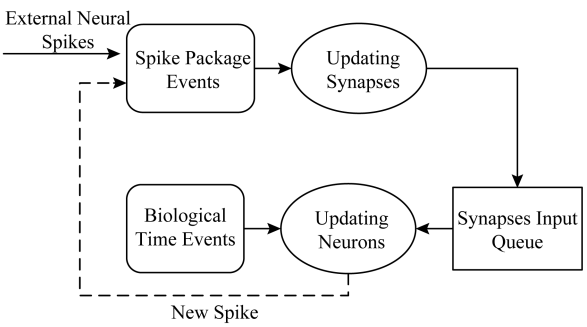


Fig. 6 Time-driven and event-driven simulation mode[32]

图 6 时间驱动和事件驱动的模拟方式[32]

综上,通过 sPyNNaker 软件系统,用户可以远程使用 SpiNNaker 虚拟机,开发和操作接口与目前主流的脉冲神经网络仿真平台类似.

4 类脑机的信息处理潜力

类脑机和大脑都是脉冲神经网络.本节先介绍采用脉冲神经网络构造任意图灵机的一种方法,它证明了脉冲神经网络的信息处理能力不低于图灵机;然后介绍脉冲神经网络如何超越人工神经网络;最后介绍噪声可以提高脉冲神经网络的性能,使得脉冲神经网络具有实现马尔可夫链蒙特卡洛(Markov chain Monte Carlo, MCMC)采样与求解约束满足 NP-hard 问题的能力.

4.1 脉冲神经网络

脉冲神经网络也被称为第三代人工神经网络[33],与前两代人工神经网络 McCulloch-Pitts-Neuron[34], Perceptron[35]不同,脉冲神经网络认为神经元脉冲发放以及脉冲之间的时间间隔也是一种重要的特性,更贴近于人脑中的真实神经元[36-37].

一个典型的生物神经元的结构如图 7 所示,主要包括树突、胞体和轴突 3 个部分.树突收集其他神经元传来的信息并通过电流的形式将其传给胞体,胞体相当于一个中央处理器,树突传来的电流引起胞体膜电位变化,当膜电位超过一定阈值时,神经元将发放一个脉冲信号(称为动作电位)并通过轴突传给其他神经元.动作电位是一个幅值大约100 mV、持续时间 1~2 ms 的电脉冲[39].

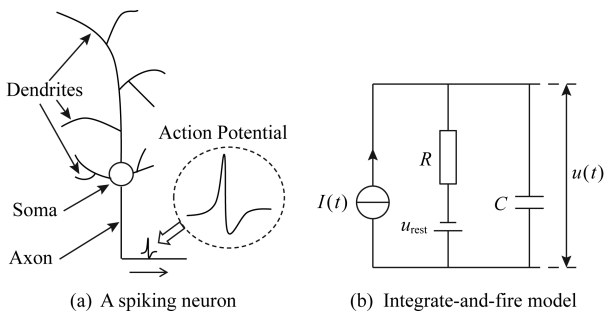


Fig. 7 A spiking neuron and integrate-and-fire model[38]

图 7 脉冲神经元示意图与积分发放模型[38]

计算神经科学家根据生物神经元的特性建立了诸多脉冲神经元模型,主要包括积分发放(integrate-and-fire)模型[40-41]、Hodgkin-Huxley 模型[3,42-44]、Izhikevich 模型[45-46]和脉冲响应(spike response)模型[47-48].这些模型以不同的精度描述了生物神经元产生动作电位的动态过程.

常用的积分发放模型最简单如图 7 所示,它将神经元的膜表示为一个电容器 C,当神经元接受输入电流时,神经元的膜电位 $u(t)$可以表示为

$$\tau \frac{\mathrm{d}u(t)}{\mathrm{d}t} = -(u(t)-u_{\mathrm{rest}})+RI(t), \quad (1)$$

其中,R,$I(t)$,u_{rest}分别表示神经元的电阻、输入电流和静息电位.当时刻 t 神经元的膜电位 $u(t)$超过一个阈值 θ 时,神经元将发放一个脉冲且膜电位复位到 $u_r<\theta$,即:

$$\lim_{\delta\to 0^+} u(t+\delta)=u_r. \quad (2)$$

前神经元的轴突与后神经元的树突相互接触之处叫做突触,它影响着神经元之间的交换信息[6-9].当突触前神经元发放一个脉冲时,将通过突触在后突触神经元的树突上产生一个电势变化,也叫做突触后电位(postsynaptic potential, PSP).突触后电位的取值可正可负,其中正的电位叫做兴奋性突触后电位(excitatory postsynaptic potential, EPSP),

负的电位称为抑制性突触后电位(inhibitory post-synaptic potential, IPSP).

单个神经元的计算能力有限,当一群神经聚集在一起构成脉冲神经网络时可以实现复杂的计算.一个脉冲神经网络可以被定义为一个图 $G=(V,E)$,其中节点 V 表示神经元的集合,边 $E\subset V\times V$ 表示突触的集合.不同脉冲神经网络可以用不同的图结构表示.

4.2 脉冲神经网络的能力不低于图灵机

本节证明利用脉冲神经网络发放脉冲之间的相位差就可以实现图灵机.

图灵机的基本思想就是用机器来模拟人们用纸笔进行数学计算的过程[49],它包括一条无限长的纸带(多条纸带为推广情况)、一个读写头、一个状态寄存器和一套控制程序指令.其中,纸带上包含一个个连续的存储格子,每个格子存储一个数字或者符号;读写头可以在纸带上移动,并可以读取纸带上的内容或者写入新的内容;状态寄存器用于存储机器当前所处的状态且机器状态数量有限;控制程序指令可以根据机器当前所处状态以及当前读写头所指格子上的数字或符号来确定读写头的移动方向(左移一格或者右移一格).理论证明,图灵机可以模拟人类所能进行的任何计算过程[50-51].

Maass[52]证明了对于任意给定的 $d\in N$,都存在一个有限规模的脉冲神经网络 $N_{\mathrm{TM}}(d)$,它可以实时模拟任意的包含 d 条无限长纸带的图灵机.主要证明包括 3 个步骤:

1) 构建几种局部脉冲神经网络,分别实现延时器、信号抑制器、振荡器、起搏器、同步器、脉冲相位大小比较器、布尔阈值电路以及相位乘法运算(相位乘以一个固定常数).

2) 证明有限规模的脉冲神经网络可以实现堆栈的功能,从而可以缓存数据.具体来说,堆栈中一串二值序列 $\langle b_1,b_2,\cdots,b_l\rangle\in\{0,1\}^*$ 可以表示为振荡器 O_S 的相位差:

$$\varphi_S=\sum_{i=1}^{l}b_i\times 2^{-i-c},\tag{3}$$

其中,φ_S 表示振荡器 O_S 与具有相同周期的一个起搏器的相位差;参数 c 用于控制相位差的大小,保证相位差小于振荡器的震荡周期,在此基础上证明入栈(push)和出栈(pop)指令可以由式(1)中定义的基本网络实现.

3) 根据文献[53],任意的包含 d 条无限长的纸带的图灵机可以类似地包含 $2d$ 个堆栈的类似的图灵机实现,其中每条纸带读写头所指处的左右 2 个部分分别用一个堆栈来表示,因此图灵机的计算可以转化为入栈和出栈的操作;再结合文献[54],可以进一步证明任意图灵机的计算过程可以通过有限个布尔阈值电路来模拟,而脉冲神经网络又可以实现布尔阈值电路,因此脉冲神经网络可以模拟图灵机的计算过程.

Maass[52]指出脉冲神经网络的能力优于图灵机,主要原因有 2 点:1)相比于图灵机,脉冲神经网络的输入输出可以为任意的实数;2)图灵机的基本操作只能作用于有限的位数,而脉冲神经网络的序列存储于相位 φ_S 中,因此基本操作可以直接改变整个序列.

4.3 噪声可以提高脉冲神经网络的性能

4.1 节中介绍的积分发放等脉冲神经元模型都是确定性模型,事实上单个神经元的离子通道门控[55]、神经递质的突触释放[56]、皮层细胞反应变异性[57]和人脑的认知活动[58]都具有随机性.Maass[59]指出噪声可以作为脉冲神经网络计算和学习的资源,可以提高脉冲神经网络的计算性能,下面分别介绍相关工作.

Gerstner 等人[48]认为噪声存在于神经元发放的阈值上,这样神经元的膜电位取任何值时神经元都可以发放,膜电位越大时发放概率越高,这样的模型也叫做随机脉冲响应模型.在此基础上 Buesing 等人[60]将神经元的发放过程理解为一个采样过程,他们证明了一个包含 K 个相互连接神经元 $z_1,z_2,\cdots,z_K$ 的脉冲神经网络可以表示一个概率分布 $p(x_1,x_2,\cdots,x_K)$,进一步若神经元的发放概率与随机变量 $x_1,x_2,\cdots,x_K$ 的后验概率满足条件(神经适应条件):

$$p(z_i(t)=1)=\frac{1}{\tau}\frac{p(x_i=1|x_{\setminus i})}{p(x_i=0|x_{\setminus i})},\tag{4}$$

则神经元的发放活动等价于 MCMC 采样.其中,$p(z_i(t)=1)$表示神经元 z_i 在时刻 t 发放一个脉冲的概率;$x_{\setminus i}$表示除了 x_i 之外的其他随机变量;τ 表示神经元发放之后的抑制期的时长.基于此结论,Buesing 等人[60]证明了脉冲神经网络可以实现边缘概率推理,他们证明了如果概率分布服从玻尔兹曼分布,则神经适应条件可以由脉冲神经网络中神经元的连接自然实现.当网络的动态性收敛时,脉冲神经元可以看作是在对平稳分布(目标分布)进行采样,统计一段时间内神经元发放时间占总时间的比例即为边缘概率.Pecevski 等人[61]指出文献[60]的研究只适用于二值随机变量,提出了 3 种方法将以上结果推广到一般图模型:1)证明通过增加辅助

变量可以将任意分布转化为玻尔兹曼分布;2)利用马尔可夫毯来扩展神经适应条件;3)利用因式分解来扩展神经适应条件.Probst 等人[62]将这 3 种方法推广到积分发放神经元模型,证明了基于电导的积分发放神经元可以实现 MCMC 采样与边缘推理.Habenschuss 等人[63]研究了脉冲神经网络采样推理的收敛速度,并证明脉冲神经网络所表示的概率分布将以指数速度收敛到平稳分布.

噪声还可以存在于神经元的突触上,Kappel 等人[64-65]发现如果在突触上叠加符合维纳过程的随机噪声,突触参数的动态性可以实现 Langvein 采样.据此他们提出了突触采样学习框架,并证明了整个网络参数所表示的分布将收敛于一个平稳分布,该框架不仅可以实现脉冲神经网络的学习,而且解释了脉冲神经网络持续重新布线的原因.Yu 等人[66]提出哈密顿突触采样学习框架,揭示了实现突触可塑性的重要分子 CaMKII 加速脉冲神经网络学习的计算机理.Kappel 等人[67]进一步将突触采样框架应用到奖励学习问题中,解释了多巴胺、STDP 和噪声时人脑强化学习的基础.

此外 Jonke 等人[68]证明了包含噪声的脉冲神经网络具有求解 NP-hard 约束满足问题的能力.其主要思想是基于机,一方面 Hopfield 等人[69]和 Aarts 等人[70]已证明玻尔兹曼机可以求解 NP-hard 约束满足问题;另一方面包含噪声的脉冲神经网络可以模拟任意的玻尔兹曼机,因此可以用脉冲神经网络求解 NP-hard 约束满足问题.Jonke 等人[68]还发现了相比于人工神经网络,脉冲神经网络在求解约束满足问题时具有更快的求解速度.

5 总结与展望

经典计算机的理论基础是图灵 1936 年奠定的,图灵机的理论边界那个时刻就已经明确.冯·诺依曼体系结构是图灵机的一种物理实现模型,采用这种体系结构的经典计算机能力的理论边界当然受限于图灵机模型.

神经网络是人工智能三大流派之一,从智能实现载体层次"自底向上"地开展研究,现在看来是构筑机器智能物理基础的最主要的可行路线.大规模神经网络的复杂结构和异步通信机制迥异于冯·诺依曼体系结构,在经典计算机上进行神经信息处理的功耗也越来越难以承受,发展面向神经网络的体系结构,对人工智能还是一般意义上的信息处理都是必由之路.

目前广泛应用的人工神经网络与生物神经网络相比,还过于简化.模拟动物大脑和人脑的精细解析有望在 20 年内逐步完成,这将成为未来神经网络体系结构的基本蓝图,基于这一蓝图研制的类脑机,将成为实现更强人工智能乃至通用人工智能的物理平台.

类脑机的思想在计算机发明之前就提出了,研究开发实践也已经进行了 30 多年,多台类脑系统已经上线运行,其中 SpiNNaker 专注于类脑系统的体系结构研究,提出了一种行之有效的类脑方案.

以 SpiNNaker 为代表的类脑机采用传统计算硬件和软件实现脉冲神经网络,因此没超出经典图灵机的范畴.随着神经形态器件的发展,未来 20 年,有望研制出逼近乃至超越生物脑的类脑机,硬件神经元和神经突触将具有真正的随机性,硬件的神经环路也将像生物神经网络一样具备丰富的非线性动力学行为,是否能够突破可计算性的理论边界、超越图灵机?这是一个尚待解决的重大理论问题,类脑机的研究开发和实现应该有助于这个问题的解决.

参 考 文 献

[1] Huang Tiejun, Shi Luping, Tang Huajin, et al. Research on multimedia technology 2015—Advances and trend of brain-like computing [J]. Journal of Image and Graphics, 2016, 21(11): 1411-1424 (in Chinese)
(黄铁军, 施路平, 唐华锦, 等. 多媒体技术研究:2015——类脑计算的研究进展与发展趋势[J]. 中国图象图形学报, 2016, 21(11): 1411-1424)

[2] Huang Tiejun. Imitating the brain with neurocomputer—A "new" way towards artificial general intelligence [J]. International Journal of Automation and Computing, 2017, 14(5): 520-531

[3] Hodgkin A L, Huxley A. A quantitative description of membrane current and its application to conduction and excitation in nerve [J]. The Journal of Physiology, 1952, 117(4): 500-544

[4] Hebb D O. The Organization of Behaviour: A Neuropsychological Theory [M]. New York: Science Editions, 1949

[5] Tsodyks M, Pawelzik K, Markram H. Neural networks with dynamic synapses [J]. Neural Computation, 1998, 10(4): 821-835

[6] Bi Guoqiang, Poo Muming. Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type [J]. Journal of Neuroscience, 1998, 18(24): 10464-10472

[7] Bi Guoqiang, Poo Muming. Distributed synaptic modification in neural networks induced by patterned stimulation [J]. Nature, 1999, 401(6755): 792-796

[8] Song Sen, Miller K D, Abbott L F. Competitive Hebbian learning through spike-timing-dependent synaptic plasticity [J]. Nature Neuroscience, 2000, 3(9): 919-926

[9] Song Sen, Abbott L F. Cortical development and remapping through spike timing-dependent plasticity [J]. Neuron, 2001, 32(2): 339-350

[10] Vogelstein J T, Amunts K, Andreou A, et al. Grand challenges for global brain sciences [J]. arXiv preprint arXiv: 1608.06548, 2016

[11] Gao Ruixuan, Asano S M, Upadhyayula S, et al. Cortical column and whole-brain imaging with molecular contrast and nanoscale resolution [J]. Science, 2019, 363 (6424): eaau8302

[12] Hodges A, Turing A. The Enigma [M]. London: Vintage, 1992

[13] Turing A. Computing machinery and intelligence—AM Turing [J]. Mind, 1950, 59(236): 433-460

[14] Neumann J V. The computer and the brain [J]. Annals of the History of Computing, 1958, 11(3): 161-163

[15] Reeke G N, Sporns O, Edelman G M. Synthetic neural modeling: The 'Darwin' series of recognition automata [J]. Proceedings of the IEEE, 1990, 78(9): 1498-1530

[16] Edelman G M. Learning in and from brain-based devices [J]. Science, 2007, 318(5853): 1103-1105

[17] Izhikevich E M, Edelman G M. Large-scale model of mammalian thalamocortical systems [J]. Proceedings of the National Academy of Sciences, 2008, 105(9): 3593-3598

[18] Mead C. Analog VLSI and Neural Systems [M]. Reading, MA: Addison-Wesley, 1989

[19] Mead C. Neuromorphic electronic systems [J]. Proceedings of the IEEE, 1990, 78(10): 1629-1636

[20] Mead C, Ismail M. Analog VLSI Implementation of Neural Systems [M]. Berlin: Springer, 1989

[21] Benjamin B V, Gao P, McQuinn E, et al. Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations [J]. Proceedings of the IEEE, 2014, 102(5): 699-716

[22] Markram H. The blue brain project [J]. Nature Reviews Neuroscience, 2006, 7(2): 153-160

[23] Modha D S, Ananthanarayanan R, Esser S K, et al. Cognitive computing [J]. Communications of the ACM, 2011, 54(8): 62-71

[24] Amunts K. Human brain project of European Union [EB/OL].[2016-07-25]. http://www.humanbrainproject.eu/ (Amunts K. 欧盟人类大脑计划[EB/OL].[2016-07-25]. http://www.humanbrainproject.eu/)

[25] Merolla P A, Arthur J V, Alvarez-Icaza R, et al. A million spiking-neuron integrated circuit with a scalable communication network and interface [J]. Science, 2014, 345(6197): 668-673

[26] Furber S B, Galluppi F, Temple S, et al. The SpiNNaker project [J]. Proceedings of the IEEE, 2014, 102(5): 652-665

[27] Brown A D, Furber S B, Reeve J S, et al. SpiNNaker—programming model [J]. IEEE Transactions on Computers, 2015, 64(6): 1769-1782

[28] Schemmel J, Briiderle D, Griibl A, et al. A wafer-scale neuromorphic hardware system for large-scale neural modeling [C] //Proc of 2010 IEEE Int Symp on Circuits and Systems. Piscataway, NJ: IEEE, 2010: 1947-1950

[29] Scholze S, Eisenreich H, Höppner S, et al. A 32 GBit/s communication SoC for a waferscale neuromorphic system [J]. INTEGRATION, the VLSI Journal, 2012, 45(1): 61-75

[30] Meier K. A mixed-signal universal neuromorphic computing system [C] //Proc of 2015 IEEE Int Electron Devices Meeting (IEDM). Piscataway, NJ: IEEE, 2015: 4.6.1-4.6.4

[31] Yan Aoshuang. Special edition of China brain project [N]. People's Daily. 2015-11-17 (in Chinese)
(闫傲霜. 北京脑计划专版[N]. 人民日报. 2015-11-17)

[32] Furber S B, Lester D R, Plana L A, et al. Overview of the SpiNNaker system architecture [J]. IEEE Transactions on Computers, 2012, 62(12): 2454-2467

[33] Maass W. Networks of spiking neurons: The third generation of neural network models [J]. Neural Networks, 1997, 10 (9): 1659-1671

[34] McCulloch W S, Pitts W. A logical calculus of the ideas immanent in nervous activity [J]. Bulletin of Mathematical Biophysics, 1943, 5(4): 115-133

[35] Rosenblatt F. Principles of neurodynamics: Perceptrons and the theory of brain mechanisms [R]. Washington, DC: Spartan Books, 1961

[36] O'Reilly R C, Munakata Y. Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain [M]. Cambridge, MA: MIT Press, 2000

[37] Izhikevich E M. Which model to use for cortical spiking neurons? [J]. IEEE Transactions on Neural Networks, 2004, 15(5): 1063-1070

[38] Yu Zhaofei. Inference and learning in spiking neural networks [D]. Beijing: Tsinghua University, 2017 (in Chinese)
(余肇飞. 脉冲神经网络的推理与学习问题研究[D]. 北京: 清华大学, 2017)

[39] Toledo-Rodriguez M, Blumenfeld B, Wu Caizhi, et al. Correlation maps allow neuronal electrical properties to be predicted from single-cell gene expression profiles in rat neocortex [J]. Cerebral Cortex, 2004, 14(12): 1310-1327

[40] Knight B W. Dynamics of encoding in a population of neurons [J]. The Journal of General Physiology, 1972, 59(6): 734-766

[41] Abbott L F. Lapicque's introduction of the integrate-and-fire model neuron (1907)[J]. Brain Research Bulletin, 1999, 50 (5/6): 303-304

[42] Hodgkin A L, Huxley A F, Katz B. Measurement of current-voltage relations in the membrane of the giant axon of Loligo [J]. The Journal of Physiology, 1952, 116(4): 424-448

[43] Hodgkin A L, Huxley A F. Currents carried by sodium and potassium ions through the membrane of the giant axon of Loligo [J]. The Journal of Physiology, 1952, 116(4): 449-472

[44] Hodgkin A L, Huxley A F. The components of membrane conductance in the giant axon of Loligo [J]. The Journal of Physiology, 1952, 116(4): 473-496

[45] Izhikevich E M. Neural excitability, spiking and bursting [J]. International Journal of Bifurcation and Chaos, 2000, 10(6): 1171-1266

[46] Izhikevich E M. Simple model of spiking neurons [J]. IEEE Transactions on Neural Networks, 2003, 14(6): 1569-1572

[47] Kistler W M, Gerstner W, Hemmen J L. Reduction of the Hodgkin-Huxley equations to a single-variable threshold model [J]. Neural Computation, 1997, 9(5): 1015-1045

[48] Gerstner W, Kistler W M, Naud R, et al. Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition [M]. Cambridge, UK: Cambridge University Press, 2014

[49] Turing A M. On computable numbers, with an application to the Entscheidungsproblem [J]. Proceedings of the London Mathematical Society, 1937, 2(1): 230-265

[50] Schacter D L. Searching for Memory: The Brain, the Mind, and the Past [M]. New York: Basic Books, 1996

[51] Sipser M. Introduction to the Theory of Computation [M]. Boston: Thomson Course Technology, 2006

[52] Maass W. Lower bounds for the computational power of networks of spiking neurons [J]. Neural Computation, 1996, 8(1): 1-40

[53] Hopcroft J E. Introduction to Automata Theory, Languages, and Computation [M]. Chennai, India: Pearson Education India, 2008

[54] Horne B G, Hush D R. Bounds on the complexity of recurrent neural network implementations of finite state machines [J]. Neural Networks, 1996, 9(2): 359-366

[55] Cannon R C, O'Donnell C, Nolan M F. Stochastic ion channel gating in dendritic neurons: Morphology dependence and probabilistic synaptic activation of dendritic spikes [J]. PLoS Computational Biology, 2010, 6(8): e1000886

[56] Flight M H. Synaptic transmission: On the probability of release [J]. Nature Reviews Neuroscience, 2008, 9(10): 736-737

[57] Azouz R, Gray C M. Cellular mechanisms contributing to response variability of cortical neurons in vivo [J]. Journal of Neuroscience, 1999, 19(6): 2209-2223

[58] Brascamp J W, Van Ee R, Noest A J, et al. The time course of binocular rivalry reveals a fundamental role of noise [J]. Journal of Vision, 2006, 6(11): 1244-1256

[59] Maass W. Noise as a resource for computation and learning in networks of spiking neurons [J]. Proceedings of the IEEE, 2014, 102(5): 860-880

[60] Buesing L, Bill J, Nessler B, et al. Neural dynamics as sampling: A model for stochastic computation in recurrent networks of spiking neurons [J]. PLoS Computational Biology, 2011, 7(11): e1002211

[61] Pecevski D, Buesing L, Maass W. Probabilistic inference in general graphical models through sampling in stochastic networks of spiking neurons [J]. PLoS Computational Biology, 2011, 7(12): e1002294

[62] Probst D, Petrovici M A, Bytschok I, et al. Probabilistic inference in discrete spaces can be implemented into networks of LIF neurons [J]. Frontiers in Computational Neuroscience, 2015, 9: 1-11

[63] Habenschuss S, Jonke Z, Maass W. Stochastic computations in cortical microcircuit models [J]. PLoS Computational Biology, 2013, 9(11): e1003311

[64] Kappel D, Habenschuss S, Legenstein R, et al. Synaptic sampling: A Bayesian approach to neural network plasticity and rewiring [C] //Proc of Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2015: 370-378

[65] Kappel D, Habenschuss S, Legenstein R, et al. Network plasticity as Bayesian inference [J]. PLoS Computational Biology, 2015, 11(11): e1004485

[66] Yu Zhaofei, Kappel D, Legenstein R, et al. CaMKII activation supports reward-based neural network optimization through Hamiltonian sampling [J]. arXiv preprint arXiv: 1606.00157, 2016

[67] Kappel D, Legenstein R, Habenschuss S, et al. A dynamic connectome supports the emergence of stable computational function of neural circuits through reward-based learning [J]. eNeuro, 2018, 5(2): ENEURO. 0301-17.2018

[68] Jonke Z, Habenschuss S, Maass W. Solving constraint satisfaction problems with networks of spiking neurons [J]. Frontiers in Neuroscience, 2016, 10: 1-16

[69] Hopfield J J, Tank D W. Computing with neural circuits: A model [J]. Science, 1986, 233(4764): 625-633

[70] Aarts E, Korst J. Simulated Annealing and Boltzmann Machines [M]. New York: John Wiley, 1989

Huang Tiejun, born in 1970. PhD. Professor and PhD supervisor. Member of CCF, ACM and IEEE. His main research interests include visual information processing and neuromorphic computing.

Yu Zhaofei, born in 1990. PhD. Member of IEEE. His main research interests include brain-like computing and machine learning.

Liu Yijun, born in 1977. PhD and professor. His main research interests include neuromorphic computing, computer architecture and integrated circuit design.