基于类脑芯片发展的综述

21307347-陈欣宇

摘要： 人工智能技术的应用逐渐广泛，深度学习对算力的需求也拉动了AI芯片的快速发展，AI芯片架构也经历了多次进化，发展出两种技术路径：一种采用传统冯·诺依曼架构，以图形处理器(GPU)、可编程逻辑阵列(FPGA)、专用集成电路(ASIC)为代表，仍然是当今市场的主流，另一种采用类脑神经结构，模仿生物大脑神经网络特征构建的存算一体化架构，类脑芯片在算力、功耗等方面都具有很大的潜力，本文主要介绍的就是类脑芯片，指出类脑芯片的特点、相关技术以及研究现况，并对类脑芯片的发展方向作出总结。

1. 类脑计算的背景

作为AI芯片，类脑计算为突破人工智能计算的发展瓶颈而生，传统的AI芯片已经经过多轮的更新迭代，最初是通过通用微处理器进行人工智能计算，由于处理器与计算需求的不匹配，伴随着人工智能领域的发展，研发出AI计算的主要加速器GPU，随后对于特定的AI计算任务，出现了AI加速硬件，如FPGA和ASIC，取得了业内广泛认同，但这些处理器都离不开将计算和存储分离的冯·诺依曼架构，计算单元速度远高于存储单元，造成了计算的速度瓶颈，也就是“内存墙”现象。同时硬件发展的Dennard和Moore这两个定律面临终结，提高晶体管性能和集成度问题受阻，功耗问题成为一大挑战。而参考人脑，所有计算和存储功能都由神经元和神经突触完成，且以超低功耗处理着大量数据，类脑计算由此得到启发，通过模仿生物脉冲来传递信息，即基于脉冲神经网络(SNN)来仿真人脑构造计算系统。

1. SNN算法

SNN(Spiking Neural Network)算法，是类脑人工智能的重要核心，为解决计算存储分的冯·诺依曼架构体系能耗低的问题提供了新思路。SNN属于第三代神经网络模型，不同于人工神经网络，SNN处理的信息参考人脑中离散的脉冲，脉冲神经元(SN)是基本的构成元素，神经元与该膜电势和激活阈值有关，前一个神经元的膜电势达到激活阈值则会放出脉冲信号传递到下一个神经元，大量神经元形成网络进行系统学习。LIF模型是SNN神经元最常用的模型，计算简单，多个神经元通过突触连接构成网络，网络拓扑结构实现的复杂度和参数选择影响SNN的学习效率，目前最常用的是拓扑结构是前馈型SNN，由输入层、隐藏层和输出层构成，每层一或多个神经元，通过权值动态调整突触连接。SNN学习算法分为无监督学习算法和有监督学习算法，无监督学习算法利用脉冲时序依赖的可塑性规则(STDP)[2]反映生物学特性，在此基础上结合卷积神经网络和监督学习任务，产生了基于STDP的有监督学习算法，还有另一种有监督学习算法由ANN转化而来，将反向传播和卷积的训练方法应用于SNN算法中，或是采用将ANN训练好的参数转化到SNN中的间接训练方法，但这种方法效率远低于直接训练。SNN算法在处理离散稀疏的数据上具有很大优势，能够有效处理图像信息以及时空数据的特征。

三、类脑芯片发展现况

随着SNN算法的进步，CPU和GPU处理器也逐渐无法满足SNN的算力要求，面向SNN的类脑计算平台成为发展所需，现在的类脑计算平台可以分为数模混合平台和全数字计算平台。其中全数字计算平台能兼容软件灵活调整SNN的结构参数。国内外类脑芯片都在该领域取得一定成果。

IBM公司研发的TrueNorth处理器正是类脑芯片研究的典型代表。TrueNorth处理器的神经元参数核连接方式是高度可配置的，可在模拟器芯片上运行相同的程序。2014年IMB公司发布第二代TrueNorth芯片，采用28nm技术，含100万个神经元核2.56亿个可编程突触，每秒可执行460亿次突触运算，功耗仅为70mW。TrueNorth芯片已应用于各种复杂任务，如利用芯片帮助卫星、高空飞机高效完成机器视觉分析[5]，以及用于3D地图生成、生物医学图像分割等。

国外其他类脑芯片还有曼彻斯特大学的SpiNNaker系统和英特尔发布的Loihi处理器，SpiNNaker系统包含多达1036800个ARM和7太字节的DRAM，可模拟人脑1%的神经元数量，Loihi支持各种STDP规则，能够配置突触延迟、自适应阈值、神经元分级连接等多项参数。相对于TrueNorth处理器无法片上学习，二者都实现片上调整神经元和突触参数的功能。

国内类脑芯片也取得一定成果。达尔文类脑芯片是浙江大学与杭州电子科技大学联合研发的国内第一款类脑芯片，集成了500万个晶体管，2048仿生神经元和约400万个神经突触，能够作为协处理器提供智能算法的硬件加速，减少对云服务的依赖。达尔文处理器具有高度可配置性，可以满足实际任务需要，但只实现了单芯片，规模较小。文献[3]基于达尔文芯片构建了一个脉冲神经网络训练仿真平台，为脉冲神经网络的训练仿真提供了便捷平台。

清华大学施路平团队研发了类脑芯片“天机芯”，包含约4万个神经元和1000万突触[6]，是全球第一款异构融合的神经形态芯片，即同时支持卷积神经网络和脉冲神经网络等多种神经网络，允许SNN和ANN之间无缝通信，解决了计算型和类脑型算法在硬件上不兼容的问题，推动了SNN的应用发展，该团队已将天机芯用于构建语音命令识别的SNN、图像处理和目标检测的CNN以及自然语言识别的LSTM等多种模型。国内类脑芯片企业北京灵汐科技有限公司还在天机芯基础上推出了HM100模组[1]，集成25万个神经元和2500万个突触，支持各类神经网络的实现和加速。

1. 忆阻器

以上提到的类脑芯片都基于CMOS的制造工艺，然而在模仿人脑神经结构中，将网络权重存储和数据权重相乘在同一位置实现，CMOS工艺实现需要较多的晶体管，导致较大的芯片面积和功耗，同时摩尔定律的延续受到物理约束，晶体管的集成度的发展遇到阻碍。在能效与集成度的制约下，新型电子元件的探索成了研究的新方向。

忆阻器可以通过外加电压的调制来改变电导值，与突触神经元的激励改变权重的特性相似，能够模拟神经突触的可塑性，且忆阻器是非易失的，能够存储突触连接的权重。这种特性能够解决传统计算机的冯·诺依曼存算分离的问题。忆阻器用于神经网络计算以阵列的形式，将权重保存为阻值，输入量化为电压，电流形式输出，计算的并行度高，随着忆阻器阵列的工艺不断优化，性能也逐渐提升，可实现信号处理、图像压缩、卷积滤波于增强学习等处理[7]，对比传统处理器具有更佳的能耗优势。基于忆阻器的类脑计算芯片架构也是主要研究方向之一，文献[4]对类脑忆阻器阵列的等效电路、尖峰脉冲信号序列进行建模和分析，并提出改进型无监督STDP规则，实现算法涉及优化。许多研究对基于多个忆阻器阵列实现复杂类脑计算进行了仿真探索，证明了忆阻类脑计算相对于传统AI芯片的优势，但仍需通过流片进行验证。有大量研究尝试将忆阻器阵列同CMOS电路进行集成，实现复杂神经网络，开发出了很多全功能忆阻类脑芯片，虽然大部分芯片的可扩展性不高，只针对一类神经网络，但也测试了全功能忆阻类脑芯片的计算功能，为集成更复杂的类脑芯片打下了基础。

忆阻器相比传统CMOS工艺有着很大潜力，适合于脉冲神经网络结构，但忆阻器工艺仍不成熟，离实用化仍有一段距离，面临着阻态不稳定、架构不成熟和外围电路复杂等问题，其生成工艺仍需要投入很大研究分析以及开发工具的支持，才能使忆阻类脑芯片慢慢走向实际应用。

1. 总结与展望

总的来看，类脑芯片采用神经拟态工程，模拟生物脑的运作特点，实现类脑的数据处理，使其能够解决复杂的计算任务，并满足低能耗的需求，具有超过GPU的计算潜力。但类脑计算由于涉及到多个学科领域，需要神经科学提供类脑计算模型，材料科学研究适用于类脑计算的新材料，类脑芯片在许多领域的探索都处于起步阶段，发展速度受到一定限制，需要充分调用神经科学、计算机科学和社会学等各领域学科交叉，逆向工程推进类脑芯片发展。以及加强忆阻器等关键材料的研究工作，类脑芯片将是人工智能芯片进化的重要一环，如今中国AI芯片产业在关键产业链上仍存在短板，但在良好的政策支持下，相信中国芯片产业的发展前景是广阔的。

参考文献：

[1]马超,辛增献,马亮等.智能信号处理平台技术综述[J].制导与引信,2022,43(04):17-23.

[2]张慧港,徐桂芝,郭嘉荣等.类脑脉冲神经网络及其神经形态芯片研究综述[J].生物医学工程学杂志,2021,38(05):986-994+1002.

[3]许晓强. 面向“达尔文”芯片的脉冲神经网平台设计与实现[D].浙江大学,2018.

[4]叶沁桐.基于忆阻器的LSTM神经网络存内计算架构设计[J].集成电路应用,2023,40(10):18-21.DOI:10.19339/j.issn.1674-2583.2023.10.007.

[5]毛磊,姚保寅,黄旭辉,王智斌.类脑计算芯片技术发展及军事应用浅析[J].军事文摘,2021,(13):57-61.

[6]赵荣杰,房超,许蔓舒.人工智能芯片产业发展现状及展望[J].经济导刊,2022(11):75-81.

[7]陈长林,骆畅航,刘森等.忆阻器类脑计算芯片研究现状综述[J].国防科技大学学报,2023,45(01):1-14.