光电智能计算*

方 璐[1,3] 吴嘉敏[2,3] 戴琼海[2,3,†]
(1 清华大学电子系 北京 100084)
(2 清华大学自动化系 北京 100084)
(3 北京信息科学与技术国家研究中心 北京 100084)

2023－12－04收到
† email: qhdai@tsinghua.edu.cn
DOI: 10.7693/wl20231205

1 引言

当前，人工智能技术的复兴正引领着新一代信息技术迅猛发展，由电子驱动的计算处理器在过去十年中发生了巨大的变化，从通用中央处理器(CPU)到定制计算平台，例如GPU、FPGA和ASIC，以满足对计算资源无处不在的持续增长的需求。这些硅计算硬件平台的进步催生了更大规模的训练和更复杂的模型，极大地促进了人工智能(AI)的复兴。我们见证了各种神经计算架构，例如卷积神经网络(CNN)、递归神经网络(RNN)、脉冲神经网络(SNN)等，在诸多领域的广泛应用。

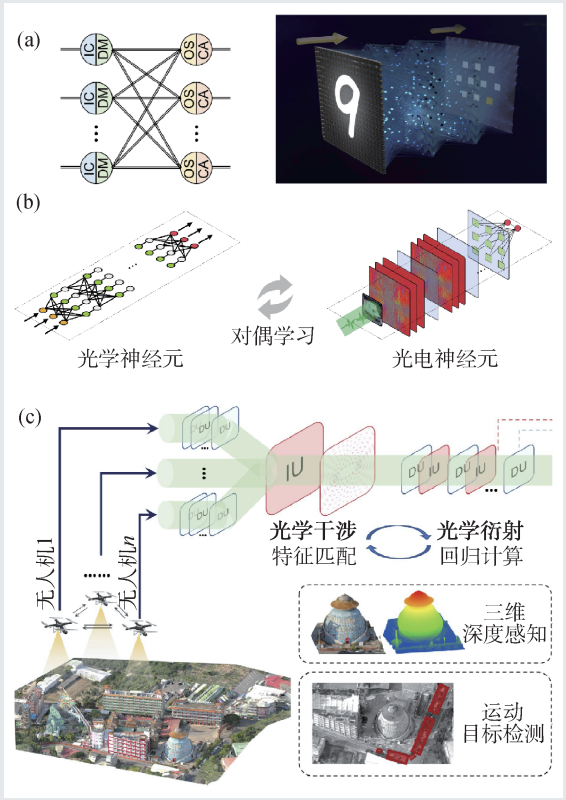

图1 大规模多通道光电智能计算架构和训练方法 (a)可重构智能计算处理器；(b)大规模光学神经网络训练；(c)多通道光电神经网络示意图

然而传统电子计算机的架构和性能的发展趋势已经无法满足新一代信息技术发展对计算资源的需求。随着先进光刻工艺的不断发展，晶体管尺寸已经缩小到10 nm以下，逐渐逼近原子尺寸，这使得芯片的加工难度以及加工成本呈指数式上升。与此同时，随着晶体管密度的增加，趋势明显的漏电流效应加剧了芯片热功耗，对系统整体散热能力的需求也不断上升，已经开始成为限制晶体管密度的另一瓶颈。故而，无论是在硬件实现还是计算架构上，都使得预测晶体管制程的摩尔定律难以维系，新型智能计算架构与芯片研究迫在眉睫。

光具有物理空间最快的传播速度以及多维度(时间、空间、光谱等)的优势，这些特性使得光计算成为构建下一代高性能计算的理想范式之一。受益于光计算的颠覆性优势(高带宽、高并行、低功耗)，相比电子计算，光计算在理论上有望提升6个数量级的能量效率、3个数量级的计算速度。针对如何实现光计算，国际上已经有初步的研究[1, 2]，一些代表性的技术包括：基于片上光学干涉仪网络实现任意矩阵变换[3]，基于谐振环和谐振腔进行可编程光计算[4]，基于衍射连接实现全光神经网络[5]，基于相变材料实现存内光计算[6]等。然而，现阶段的光计算仍然面临算力不足、动态计算困难、训练效率低下等问题，如何实现大规模、可重构、低功耗的光电计算芯片并支撑人工智能应用仍然面临原理架构、智能算法、集

* 国家自然科学基金(批准号：62125106；62088102)资助项目，科技部重大项目(批准号：2021ZD0109901)

成工艺等诸多难题。

2 光电智能计算架构和芯片研究

2.1 光电智能计算架构

针对光电智能计算面临的规模与重构难题，清华大学研究团队提出了可重构衍射智能计算架构，构建了可重构衍射智能计算处理器(DPU)(图1(a))[7]。DPU对光学衍射物理现象进行建模，通过大规模的光学互联，构建高复杂度的光学神经网络(图1(b))。此外，DPU充分挖掘了光的波粒二象性，控制光波传播的波前分布，实现光神经网络权重的调整，采用光电效应来实现人工神经元，解决大规模光电非线性激活函数这一理论难题。通过高通量可编程的光电器件结合电子计算的灵活特性，实现了高速数据调控以及大规模网络结构和参数的编程。DPU计算架构中，光计算模块几乎承担了所有的计算操作。因此，运行同样的神经网络，光电计算系统与特斯拉V100图形处理器(GPU)相比，计算速度提高了8倍，系统能效提升超过一个数量级，核心模块计算能效可以提升4个数量级。

研究团队进一步对光学干涉与衍射进行联合建模，提出了多通道光电神经网络的新架构Monet (multi-channel optical neural NETworks)[8]，将多个光学通道的光场信息进行融合计算，实现了基于光电智能计算的高维光场信息调制解耦(图1(c))。其中，编码投影干涉计算单元(IU)，通过相位和偏振的编码调制以及通道间的光学干涉，实现特征匹配、加权求和等多通道光学基本计算。IU和衍射计算单元(DU)的交替级联，实现了光场信息的多通道可重构智能计算。Monet架构突破了现有光电神经网络结构简单、通道受限等瓶颈，为构建大规模光电神经网络、探索复杂光场智能感算提供了理论与架构支撑。

目前光电智能计算在高速图像处理方面有突出表现，但现有架构难以挖掘高速动态光场的时间维度特性，动态计算受制于电子内存读写的瓶颈，难以满足面向超快动态现象开展实时智能分析的现实需求。我们提出了空时域智能光计算架构[9]，刻画多维光场传播模型，建立空时域光计算表征，在空间和时序维度上同时完成连续光计算(图2(a))。我们还提出了空间复用和光谱复用的智能计算模型(图2(b))，匹配空时域光计算维度，建立时序矩阵乘加计算模型，实现了三维空时域智能光计算。空时域光计算的空间和时序计算操作均在光学模拟域完成，突破了数字内存读写的掣肘，将动态机器视觉处理的速度提升了3个数量级(达到纳秒量级)。

现有光电神经网络学习架构仅能支撑小规模训练，其网络容量和特征捕获能力不足以有效处理ImageNet等大型复杂数据集。为了解决大规模光电神经网络中优化速度慢、资源消耗高、收敛效果差等问题，研究团队提出了面向大规模光电智能计算的“光学—人工双神经元学习架构DANTE(DuAl-Neuron opTical-artificial lEarning)[10]。其中光学神经元精准建模光场计算过程，人工神经元以轻量映射函数建立跳跃连接，助力梯度传播，全局人工神经元与局部光学神经元以交替学习的机制进行迭代优化，在确保学习有效性的同时，大大降低了训练的时空复杂度，使得训练更大更深的光电神经网络成为可能。DANTE突破了大规模光电神经网络物理建模复杂、参数优化困难等桎梏，网络规模提升一至两个数量级，训练学习速度提升2个数量级。

2.2 全模拟光电智能计算芯片

在上述系列新架构的基础上，研究团队研制了国际首个全模拟光电智能计算芯片ACCEL

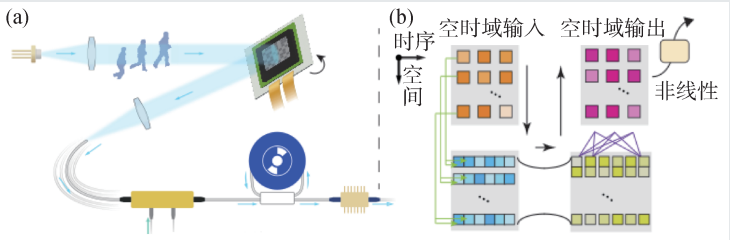

图2 空时域动态光计算 (a)空时域动态光计算系统示意图；(b)空时域动态光计算模型

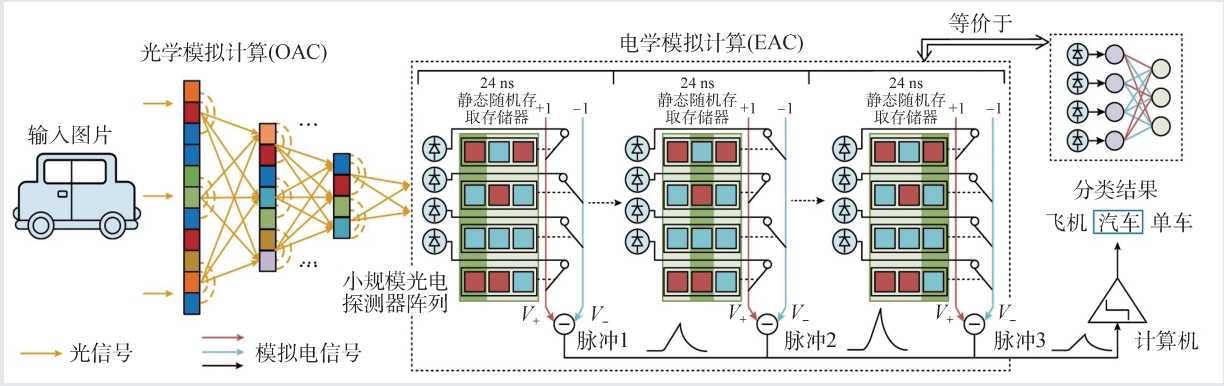

图3 全模拟光电智能计算芯片(ACCEL)工作原理图

(图3)[11, 12]，在一枚芯片上突破性地实现了大规模计算单元集成、光计算与电子信号计算的高效接口。其核心思想是通过全模拟的光电计算方式来降低对大规模光电二极管阵列和高功耗模拟数字转换器(ADC)阵列的依赖，实现光学和电子计算的高效集成。ACCEL的工作原理涉及两个主要模块，即光学模拟计算(OAC)和电子模拟计算(EAC)。OAC通过多层衍射光学计算模块，以光速提取高分辨率图像的特征，降低图像维度并减少光电转换需求。EAC包括一个32×32的光电二极管阵列，作为非线性激活器，将光学信号转换为模拟电子信号，实现类似二进制加权的全连接神经网络。ACCEL芯片以全模拟方式进行计算，适用于广泛的应用，并与数字神经网络兼容。

ACCEL通过数值模拟和实验验证，在低光条件下展现出优异的稳健性。对于输入光强的降低，ACCEL通过模拟噪声对输出进行精准校准，可以成功应对多种干扰。在可重构方面，同一OAC在不同任务中均取得了显著效果。通过OAC对多个数据集的联合训练，ACCEL在不同领域的应用中取得了出色的泛化性能，为实际工业检测等场景提供了关键的灵活性。与现有高性能芯片相比，ACCEL芯片的算力(单位时间的运算次数)提升了3000倍，系统级能效(单位能量可进行的运算数)提升了400万余倍。对于10类MNIST分类和3类ImageNet分类，ACCEL各达到9.49×10^3 TOPS/W和7.48×10^4 TOPS/W (1 TOPS/W表示在1 W功耗的情况下，处理器可以进行10^{12}次操作)的系统能效，展示了其在能效方面的优越性。ACCEL作为一种全新的光电神经网络，通过其独特的设计和卓越的性能，在人工智能硬件领域崭露头角。其在图像分类、视频判断和低光条件下的稳健性等方面的优异表现，为未来神经网络研究和应用开辟了新的前景。

3 总结

光电智能计算作为一种新兴计算范式，将为后摩尔时代的人工智能高效训练和推理带来新的契机。光子智能芯片的研究将极大促进人工智能的发展，为大规模数据的高效智能处理、大场景多对象光场智能感算、高速低功耗智能无人系统、超高速科学研究等奠定基础，具有广阔的应用前景。

参考文献

[1] Wetzstein G *et al*. Nature, 2020, 588: 39
[2] Shastri B J *et al*. Nature Photonics, 2021, 15: 102
[3] Shen Y *et al*. Nature Photonics, 2017, 11(7): 441
[4] Tait A N *et al*. Scientific Reports, 2017, 7(1): 7430
[5] Lin X *et al*. Science, 2018, 361(6406): 1004
[6] Feldmann J *et al*. Nature, 2021, 589: 52
[7] Zhou T *et al*. Nature Photonics, 2021, 15: 367
[8] Xu Z *et al*. Light: Science & Applications, 2022, 11(1): 255
[9] Zhou T *et al*. Science Advances, 2023, 9(23): eadf4391
[10] Yuan X *et al*. Nature Communications, 2023, 14(1): 7110
[11] Chen Y *et al*. Sci. Adv., 2023, 9: eadf8437
[12] Chen Y *et al*. Nature, 2023, 623: 48