

引用格式: 谢斌, 谢红. 光子芯片的智能化前景: 类脑智能算法的光学实现与潜在突破[J]. 世界科技研究与发展, 2024, 46(6): 729 – 757.

光子芯片的智能化前景: 类脑智能算法的光学实现与潜在突破*

谢斌^{1,2,3} 谢红^{*,1,2}

(1. 上海理工大学智能科技学院, 上海 200093; 2. 上海理工大学光子芯片研究院, 上海 200093;
3. 上海理工大学健康科学与工程学院, 上海 200093)

摘要: 当前, 以人工神经网络为代表的类脑智能模型主要在软件层面实现。尽管这些模型在智能化方面已经取得了重大突破, 但其发展仍然受到电子芯片物理限制的制约。随着光计算技术的发展, 基于模拟光计算的光子芯片能够在硬件层面上直接构建具有神经形态的物理计算单元, 实现高效智能处理与自适应学习。本文首先概览了数字光计算和光量子计算领域中光子芯片的相关研究进展, 指出发展基于模拟光计算的类脑光子芯片能够规避逻辑门设计的问题并模拟生物脑在计算方面的优势。通过梳理人工神经网络的演化以及相应光学实现技术, 本文针对现有类脑光子芯片在非线性元件和规模扩展方面的瓶颈, 进一步提出了以下观点: 在硬件实现层面, 需要进一步探索非线性光学元件, 构建全光非线性操作层来模拟脑的非线性特性; 在算法实现层面, 需要重点研究大脑以记忆为基础的学习认知原理, 并设计匹配光子特性的类脑智能算法, 从而突破规模限制。

关键词: 光子芯片; 光学神经网络; 光子突触; 类脑光子芯片; 类脑智能算法

DOI: 10.16507/j.issn.1006-6055.2024.11.002

The Future of Intelligent Photonic Chips: Optical Implementation of Brain-inspired Algorithms and Potential Breakthroughs*

XIE Bin^{1,2,3} XIE Hong^{*,1,2}

(1. School of Artificial Intelligence Science and Technology, University of Shanghai for Science and Technology, Shanghai 200093, China; 2. Institute of Photonic Chips, University of Shanghai for Science and Technology, Shanghai 200093, China; 3. School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: Currently, brain-inspired models, represented by artificial neural networks, are primarily implemented at the software level. Although these models have achieved significant breakthroughs in terms of intelligence, their development is still constrained by the physical limitations of electronic chips. With the advancement of optical computing technology, photonic chips based on analog optical computing can directly construct physical neuromorphic computing units at the

* 国家自然科学基金“视皮层发育关键期可塑性的突触环路调控机制”(32130043), 上海市教委“上海市类脑光子芯片前沿科学研究基地”(NO. 2)

* * E-mail: hongxie@usst.edu.cn

hardware level, enabling efficient intelligent processing and adaptive learning. This paper first provides an overview of recent research progress on photonic chips in the fields of digital optical computing and optical quantum computing, highlighting that developing brain-inspired photonic chips based on analog optical computing can circumvent the issue associated with logic gate design and simulate the computational advantages of the brain. By reviewing the evolution of artificial neural networks and the corresponding optical implementation technologies, this paper further presents the following viewpoints regarding the bottlenecks in existing brain-inspired photonic chips, specifically in nonlinear components and scalability. In terms of hardware implementation, it is necessary to further explore nonlinear optical components and construct all-optical nonlinear operation layers to simulate the nonlinear characteristics of the brain. In terms of algorithm implementation, it is crucial to study the memory-based learning and cognitive principles of the brain to design brain-inspired intelligent algorithms that match photonic characteristics, thereby overcoming scalability limitations.

Keywords: Photonic Chip; Optical Neural Network; Photonic Synapse; Brain-inspired Photonic Chip; Brain-inspired Algorithm

近年来,以人工神经网络为代表的类脑智能模型取得了飞速的发展。其中,基于深度学习技术^[1]的大参数量模型智能化水平不断提升,在一些特定任务中的表现甚至超越了人类^[2],并且已经在图像识别^[3]、自然语言处理^[4]和自动驾驶^[5]等领域实现了工业应用。然而,这些模型大多是在软件层面上构建的。在硬件层面,模型依赖于将算法转换为二进制机器码,并通过传统电子芯片的数字计算单元进行处理。随着摩尔定律逐渐接近其物理极限,基于冯·诺依曼架构(Von Neumann Architecture)^[6]或哈佛架构(Harvard Architecture)^[7]的电子芯片面临着功耗、散热、互连延迟、制造成本和可靠性问题等方面的挑战^[8-10]。这些挑战促使行业积极探索新的计算技术,以应对类脑智能模型不断增长的算力需求。光子芯片的出现为类脑智能的发展提供了新的思路:通过利用光子的相关计算特性,光子芯片能够在硬件层面上直接构建具有神经形态的物理计算单元,从而具备模仿生物脑进行低能耗大规模并行计算的能力,实现类脑的信息处理方式。智能化光子芯片为构建类脑智能系统提供了高能效的硬件基础,并有望突破电子芯片的瓶颈,成为新一代神经网络的部署平台。在这一背景下,本文论述了光子芯片的智能化进程,并对类脑光子芯片

发展提出了建议。

1 光子芯片数据计算方式的发展概览

光子芯片的数据计算方式可以按照理论的不同分为两类:基于经典波动光学和非线性光学的光子计算以及基于量子光学的线性光学量子计算。这两种计算方式运用了光子的不同特性来实现数据的传输和处理,推动了经典计算和量子计算领域中许多变革性技术的发展。为了更深入地理解这两种计算方式在构建光子芯片过程中的应用与挑战,本文将进一步展开讨论。首先,1.1 数字光计算与光量子计算这一小节将介绍数字光和光量子实现逻辑门的原理与相关难点。随后,1.2 基于模拟光计算的类脑光子芯片这一小节将探讨模拟光计算在类脑智能和神经网络模拟中的优势,并分析其在实现高效、低能耗计算方面的潜力与实际进展。

1.1 数字光计算与光量子计算

1.1.1 数字光计算

比特(Binary digit, Bit)是经典数字计算中信息量的基本度量单位,表示二进制系统中的一个数位,可以有两种状态:0 和 1。在电子芯片中,比特的两种状态可以被描述为经典物理学上与电子相关的状态,例如电压的高低和电流的有无等;而

在基于数字光计算的光子芯片中,则往往被描述为光脉冲的有无、偏振的方向、相位的角度和光的强度等。

数字计算的本质是通过各种逻辑运算来改变比特的状态,从而实现信息的处理和转换。基于半导体导电性和掺杂特性制造而成的晶体管 (Transistor) 是电子芯片中用于控制电流并构建逻辑运算基本单元——逻辑门的基础元件。多个晶体管的组合可以实现复杂的逻辑运算操作,进而完成数据计算任务。然而,在传统的线性光学中,光在介质中传播时不会与其他光相互作用,这意味着光波彼此之间不会直接影响或改变对方的传播状态。因此,单纯依赖线性光学效应来实现逻辑运算是非常困难的,因为逻辑运算通常需要通过信号间的相互作用来产生明确的输出信号。

尽管无法完全替代传统电子芯片中逻辑门的功能,线性光学中的干涉效应仍提供了一种能够实现基本逻辑操作的途径:多束相干光波相遇时,它们可以根据彼此间的相位差在空间区域内相互叠加形成特定的干涉图样,产生具有特定强度分布的光信号,从而在一定程度上模拟逻辑运算的结果。通过设计适当的干涉结构,例如马赫-曾德尔干涉仪 (Mach-Zehnder Interferometer, MZI)^[11] 和微环谐振器 (Micro Ring Resonator, MRR)^[12],就可以利用光波的干涉特性来执行与逻辑门类似的功能。然而,这类逻辑门通常需要使用电信号驱动,不是真正意义上的全光逻辑门,且逻辑门内部电信号与光信号之间的电光转换存在损耗^[13]。

随着非线性光学的引入,逻辑门的光学实现成为了可能。目前,组成全光逻辑门的基础元件主要有五种类型:半导体光放大器 (Semiconductor Optical Amplifier, SOA)^[14-20]、高非线性光纤 (High Nonlinear Fiber, HNLF)^[21-25]、硅基光波导^[26-29]、周

期性极化铌酸锂 (Periodically Poled Lithium Niobate, PPLN) 光波导^[30-32] 以及插入非线性克尔 (Kerr) 介质的光子晶体^[33-35]。表 1 比较了各类基于非线性效应的全光逻辑门在生产中的优势和劣势。当前,在芯片应用中最具发展潜力的是基于硅基光波导的全光逻辑门,一方面是因为硅的高折射率可以使波导在微米甚至纳米尺度上实现超紧凑设计;另一方面是因为硅基光波导与互补金属氧化物半导体 (Complementary Metal-Oxide-Semiconductor, CMOS) 工艺相兼容,大规模集成优势十分显著。然而,级联多个独立的全光逻辑门并构造复杂逻辑系统的相关工作仍未大规模实现。考虑到现有电子芯片中的晶体管数量已达数亿^[36],全光逻辑门的研究与实际工业应用之间仍有很大差距。

1.1.2 光量子计算

量子计算是一种遵循量子力学规律调控量子信息单元——量子比特 (Qubits) 进行计算的新型计算模式。基于经典物理学的传统二进制计算模式只能顺序处理单一的计算路径。而在量子计算中,量子叠加 (Quantum Superposition) 效应使得量子比特在同一时间处于多个状态,从而能够实现多个计算路径的处理;并且,量子纠缠 (Quantum Entanglement) 效应使得多个量子比特之间形成了紧密关联的状态,计算时对其中一个量子比特的操作可以即时影响其他量子比特,进而实现并行处理。

线性光学量子计算系统以光子的偏振、路径、轨道角动量和空间模式等自由度为载体来构造量子比特进行量子计算。与其他量子计算系统相比,线性光学系统的优势在于:1) 光子常温下退相干时间长,因此系统可执行更多的量子操作;2) 光子可以通过光波导传输给其他设备,因此系统易于和基于经典物理学的计算系统集成。

表 1 全光逻辑门的比较

Tab. 1 Comparison of All-Optical Logic Gates

基础元件	逻辑门实现原理	优势	劣势
半导体光放大器	交叉增益调制 ^[17] 、 交叉相位调制 ^[16] 、 交叉偏振调制 ^[19] 、 四波混频 ^[20]	结构简单,单个半导体光放大器 就可实现各种逻辑门 ^[18]	存在自发辐射噪声 ^[15] ;载流子 恢复时间长,无法实现高速的信号处理 ^[14]
高非线性光纤	自相位调制 ^[21] 、 交叉相位调制 ^[22] 、 四波混频 ^[24] 、 偏振旋转 ^[25]	具备高速处理光信号的能力;与 现有的光纤网络和设备相兼容	体积大,难以集成到小型化的芯 片上 ^[23]
硅基光波导	交叉相位调制 ^[28] 、 四波混频 ^[27] 、 受激拉曼散射 ^[29]	与 CMOS 工艺相兼容,易于大规 模集成	硅材料中会发生双光子吸收和 自由载流子吸收效应,产生 损耗 ^[26]
周期性极化 铌酸锂光波导	差频 ^[32] 、 和频 ^[30] 、 倍频 ^[31]	非线性极化率高,能够实现高效 的非线性光学过程	铌酸锂的光学属性对温度敏感, 需要精确的温度控制来维持恒 定的相位匹配条件,从而保持有 效的非线性过程
光子晶体	四波混频 ^[35] 、 受激拉曼散射 ^[34]	具有周期性结构,设计简便	非线性克尔介质对工作频率高 度敏感,因而带宽较窄 ^[33]

与经典计算中的情形类似,基于量子计算的光子芯片也需要设计一系列量子逻辑门来改变量子比特的状态,从而实现量子信息的处理和转换。量子计算的基本门操作按照处理量子比特的不同可分为基于单量子比特的逻辑门和基于多量子比特的逻辑门。对于线性光学系统来说,单量子比特逻辑门的构建相对容易。以操控单光子偏振量子比特为例,可以通过四分之一波片和半波片的组合实现^[37]。但当构建多量子比特逻辑门时,由于光子间的相互作用很弱,线性光学系统只能通过后选择(Post-Selection)的方式实现概率性的多量子比特逻辑门^[38]。这意味着这些量子逻辑门在给定的输入状态下,并不是每次都能成功地执行预期的操作。因此,多量子比特逻辑门的成功率往往较低^[39],这也成为了线性光学量子计算的一个瓶颈,使得线性光学量子计算系统并不能够完全满足 DiVincenzo 标准^[40]中的所有条件并实现大规模且通用化的量子计算。

虽然通用量子计算难以实现,但光子芯片在

基于哈密顿量(Hamiltonian)映射的专用量子计算领域中实现了广泛的应用,光子芯片能够通过
对哈密顿量的精准调控实现各种量子模拟问题。例如,Tang 等^[41]通过一种三维集成光子芯片实现了对光合作用能量传输过程的量子模拟。在这项工作中,光子芯片被用来映射绿硫细菌内部色素-蛋白复合物(来自三位学者的姓名 Fenna-Matthews-Olsen,FMO)中七个菌绿素分子的哈密顿量。通过对波导结构的设计,Tang 团队在光子芯片上构建了一个与实际 FMO 复合物结构相匹配的哈密顿矩阵。这个哈密顿矩阵确保了光子芯片的量子模拟过程能够精确反映出 FMO 复合物中的耦合关系和能量传输特性。

1.2 基于模拟光计算的类脑光子芯片

在电子芯片的数字计算范式中,首要任务是实现单个逻辑门,以进行基本的逻辑运算;其次,需要级联多个独立的逻辑门,以构造更复杂的逻辑计算模块;最后,应当根据相应计算任务的需求,大规模集成不同的逻辑计算模块,以完成芯

片设计。正如 1.1 节所述,无论是经典计算框架还是量子计算框架下的光子芯片,当前试图遵循和复制电子芯片数字计算范式的研究工作都面临着逻辑门难以构造的挑战。

然而,基于模拟光计算的类脑光子芯片并不存在类似的问题。与数字光计算和量子计算相比,基于模拟光计算的类脑光子芯片可以规避逻辑门设计的问题:只需要定义神经元结构和突触权重更新法则,而无需具体定义每个神经元要实现的逻辑功能,就可通过训练的方式使得根据类脑智能算法构造的神经网络具备完成计算任务的能力。这主要源于脑的特性:1) **大脑的基本功能单元是神经元**,神经元的工作方式不同于逻辑门执行简单的逻辑操作,可接收来自多个突触的输入信号,并通过复杂的生物机制综合这些输入,以决定是否触发动作电位来传递信号;2) **大脑的计算是高度并行的**,数以亿计的神经元在同时进行信号处理,而逻辑门通常以串行或有限的并行方式执行简单的、独立的操作;3) **大脑的神经网络具有高度可塑性**,能够通过学习和经验改变神经元之间的连接强度^[42,43]。这种可塑性使得大脑能够高效地进行记忆,而逻辑门则是固定的、不可塑的,只能执行预先定义好的逻辑操作。

同时,模拟光计算中光子的相关特性也确保了类脑光子芯片能够更好地模拟生物脑在计算方面的优势——并行计算、低功耗、高存储密度和强鲁棒性。其中,光子的特性主要体现在以下 5 个方面。

1) **高速传输**。传统电子芯片主要通过电互连(Electrical Interconnection)的方式进行数据传输,即使用电导体将电子元件或电路连接起来,

以实现信号的传输。随着各种设备对算力需求的增加,芯片已不再局限于单个核心。未来,芯片将集成数百甚至数千个内核参与多线程的数据处理以提升芯片计算速度^[44,45]。在多核芯片中,内核之间的互连距离通常较长。由于受到 RC 延迟(Resistive-Capacitive Delay)^[46]的影响,电信号传播时间会随着互连距离的增加而增加,进而导致数据传输速度降低。而光子芯片中的数据传输路径是通过基于光波导的光互连(Optical Interconnection)系统搭建而成。光子作为信息载体时可在光波导中以接近光速的速度传输数据,并且光波导的损耗很低,典型值约为 1 dB/cm^[47],这意味着光信号可以在长互连距离下高速传输而不显著衰减。

2) **宽带宽**。波分复用(Wavelength Division Multiplexing, WDM)技术是一种在单根光波导上同时传输多个波长光信号的技术,其中每个波长的光信号都被视作一个数据通道。光子芯片通过波分复用技术可以叠加多个波长光信号的带宽,从而显著提高芯片总的传输带宽并减少芯片所需的 I/O 引脚数^[48]。对于多核芯片来说,芯片带宽需求将超 100 Gbps^[49],而理论上基于波分复用的数据传输方式带宽可超过 1 Tbps^[50,51]。

3) **低功耗**。随着电子元件的尺寸减小以及电路功能和复杂性的增加,电子芯片需要实现较高的金属导线密度^[52]。为了在有限的芯片空间内容纳更多的导线,必须缩小导线宽度以满足布线的需求。然而,导线宽度的缩小会导致电阻增加,进而导致功耗增加,因此金属导线密度存在瓶颈。同时,为了进行长互连距离的信号传输,电子芯片需要使用多级中继器(Repeater)^[53]重新整形和放大电信号来减少传输过程中的损耗,

这进一步增加了芯片总体的功耗。而光子芯片利用光子来传输信号,光信号在波导中传输时不会受到类似电子迁移引起的热效应和电阻损耗,因此光子芯片可以以更低的功耗进行高速数据传输。同时,光信号可以在长互连距离内高效传输而不需要进行频繁的信号放大^[54],这也为芯片节省了大量功耗。

4) **抗电磁干扰。**在电子芯片中电信号会产生电磁干扰 (Electromagnetic Interference, EMI),这种干扰可能影响周围其他电子元件的正常工作,因此需要设置额外的屏蔽层来减少干扰^[55],这不仅增加了芯片设计的复杂性和成本,还占用了宝贵的芯片空间。而在光子芯片中,光子本身不带电荷,既不会对其他元件产生电磁干扰,也不会受到其他外部干扰的影响。

5) **高存储密度。**与传统电子芯片采用存算分离的体系不同,类脑光子芯片旨在实现存算一体化。因此,芯片的计算单元不仅要实现功能集成,还必须具备高存储密度,以支持复杂的数据处理和存储需求。顾敏院士领导的上海理工大学和中科院上海光学精密机械研究所团队利用国际首创的双光束调控聚集诱导发光 (Aggregation-Induced Emission, AIE) 超分辨光存储技术,构造了单盘等效容量达 Pb 量级的三维纳米级光盘记忆体^[56]。该技术实现的存储容量与大脑的神经总数相当,为类脑光子芯片的存算一体化进程奠定了基础。围绕这一技术,顾敏院士团队进一步提出了模仿神经形态和大脑记忆印迹神经环路的光存储方案^[57]。该方案从神经科学中记忆印迹细胞的角度出发,提出了以纳米光子学技术模拟大脑记忆方式的途径,暗示了结合纳米光子学与生物学相关概念构建存算一体结构的可行性。

2 类脑智能算法的发展与光子芯片的智能化进程

2.1 人工神经网络算法的演化

M-P 模型^[58] (McCulloch-Pitts Model) 是由美国心理学家 McCulloch 和数学家 Pitts 于 1943 年所提出的一种用于模拟生物脑神经元基本工作方式的数学模型。如图 1 所示,在 M-P 模型中,神经元接收多个二进制输入,每个输入乘以权重后求和,若总和超过阈值则代表神经元被激活并触发了动作电位,否则代表神经元被抑制。

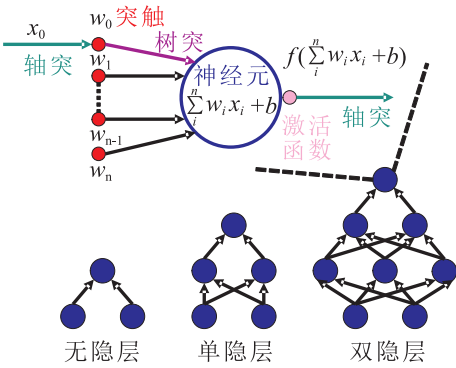


图 1 神经元的数学模型与神经网络的层级结构

Fig. 1 The Mathematical Model of a Neuron and the Hierarchical Structure of Neural Networks

M-P 模型是第一个将神经元的生物学行为形式化为数学模型,并用于模仿脑处理信息机制的尝试,它建立了后续神经网络研究的基础框架。表 2 梳理了类脑智能算法的演化进程,以 M-P 模型为基础的第一代神经网络是一种只有输入层和输出层的无隐层网络结构,被称为感知机^[59] (Perceptron),它的输入为样本的特征向量,而输出为预测的样本类别。感知机旨在求出对应于输入空间中样本划分为两类的分离超平面。与 M-P 模型预先定义突触权重的方式不同,感知机的突触权重是在求解超平面的过程中,通过极

表 2 类脑智能算法的演化

Tab. 2 Evolution of Brain-inspired Intelligent Algorithm

关键时间点	算法名称	主要贡献	存在的缺陷
1943	M-P 模型 ^[58]	首次构建了脑中神经元的数学模型,用于模拟脑处理信息的机制。	缺少突触权重更新法则的相关定义,突触权重和激活函数的阈值是预先设定且固定不变的。
1958	感知机 ^[59]	引入了前向传播和损失函数的概念,用于衡量模型性能的损失;定义了基于极小化损失函数的突触权重更新法则,用于调整无隐层神经网络的权重。	只适用于线性可分问题
1986	多层感知机 ^[60]	引入了隐藏层和非线性激活函数的概念,用于解决线性不可分问题;设计了基于反向传播的突触权重更新法则,用于调整含有隐藏层的神经网络的权重。	处理高维数据、时序数据和复杂空间关系时,存在局限性,并且容易出现过拟合和训练困难等问题。
1990	循环神经网络 ^[62]	引入了循环连接的概念,用于解决时序数据处理问题。	在捕捉长距离依赖关系时会遇到梯度消失的问题,进而无法通过梯度下降法更新参数。
1998	卷积神经网络 ^[63]	引入了感受野的概念,用于解决空间结构处理问题;提出了权重共享机制,用于减少模型参数量并提高模型泛化能力。	不适合捕捉空间特征的全局关系
2017	Transformer ^[64-67]	引入了自注意力机制,用于克服了循环神经网络和卷积神经网络的局限性;统一了自然语言处理和计算机视觉的基础模型,开启了利用大参数量模型处理多模态任务的时代。	预训练需要庞大的计算开销

小化基于误分类的损失函数而确定的。具体来说,感知机首先通过输入层到输出层的前向传播计算得到样本类别的预测值,然后将预测值与样本的实际标签作比较,并计算其中误分类的样本点到当前参数下定义的超平面的距离之和,最后将这个距离和作为损失函数并通过随机梯度下降法更新突触权重。

遗憾的是,单层感知机只能处理线性可分的问题。例如,它可以通过学习适当的权重和偏置来实现逻辑运算 AND 和 OR,但无法解决 XOR (异或)问题,因为 XOR 问题是线性不可分的。线性不可分的问题极大限制了感知机的应用,并使类脑智能算法的研究停滞了很长时间。

为了克服单层感知机的局限性,研究者们在单层感知机的输入层和输出层之间设计了多个隐藏层(Hidden Layers),并将单层感知机中线性的基于阈值的激活函数替换为了非线性的激活函数,由此提出了基于多层感知机^[60] (Multilayer

Perceptron, MLP)的第二代神经网络。如图 2 所示,通过隐藏层的多次非线性变换,即使输入数据是线性不可分的,隐藏层最终的输出也可以是线性可分的^[61],这种多层次的非线性变换使得 MLP 能够将线性不可分问题映射成线性可分问题,并完成复杂的模式识别任务。然而,由于网络结构是多层次的,MLP 无法利用单层感知机中的突触权重更新法则直接对隐藏层的权重进行更新,因为隐藏层的输出值和样本实际标签间的关系是未知的,无法作比较。对此,研究者们设

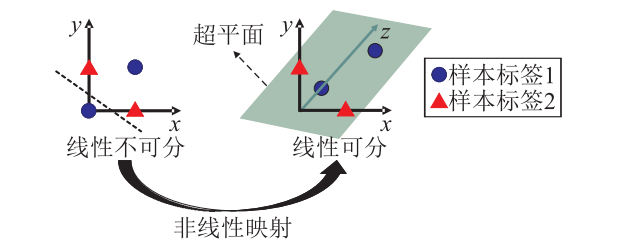


图 2 隐藏层转换线性不可分问题的图示

Fig. 2 Illustration of Linear Inseparability Transformation in Hidden Layers

计了基于反向传播 (Backpropagation, BP) 的突触权重更新法则: 根据链式求导法则, 计算损失函数相对于每个权重的梯度, 并通过梯度下降法进行多次迭代, 使得 MLP 网络能够逐渐逼近输入数据和目标输出之间的复杂非线性关系。

然而, MLP 在处理高维数据、时序数据和复杂空间关系时存在局限性。1) MLP 的每一层神经元与上一层的所有神经元都连接, 这种全连接的结构导致模型的参数量在高维数据下迅速增长, 造成了极大的计算开销。2) MLP 假设输入特征是独立的, 并且没有考虑特征之间的局部关系或相互依赖关系。而在高维数据中, 特征之间往往有复杂的关联性, MLP 因此难以有效捕捉这些关系。3) MLP 的结构设计是面向静态数据的, 即假设所有输入特征是一次性给定的, 且特征之间没有顺序依赖性。时序数据 (如时间序列、文本、语音等) 具有显著的时间依赖性, 需要模型记住和利用前后数据之间的关系。MLP 无法直接捕捉和建模这种时间依赖性, 导致它在处理时序数据时表现不佳。4) MLP 将输入数据视为扁平化的向量, 忽略了数据内部的空间结构。在处理图像这种具有空间结构的数据时, 像素之间的局部关系非常重要, 而 MLP 无法捕捉这些局部特征, 从而在图像处理任务上表现不佳。

考虑到上述 MLP 的痛点, 针对时序数据处理问题, 研究者们设计了一种基于循环连接结构的循环神经网络^[62] (Recurrent Neural Network, RNN)。RNN 通过在网络中引入循环连接, 使得每个时间步的输出不仅依赖于当前的输入, 还依赖于前一个时间步的隐藏状态。这种机制允许网络保留和利用先前时间步的信息, 从而有效地处理时序数据。然而, 当时序数据过长时, 梯度在多个时间步之后会逐渐消失, 导致 RNN 难以继

续通过基于 BP 的梯度下降法更新突触权重值。

针对空间结构处理问题, 研究者们设计了一种基于感受野的卷积神经网络 (Convolutional Neural Network, CNN)^[63]。CNN 使用卷积层来处理具有空间结构的数据。当卷积核进行局部感受野操作时, 每个卷积层只会处理输入数据中的局部信息, 这使得网络能够识别数据中的边缘、角点和纹理等局部空间特征信息。同时, 相比于全连接, 卷积操作中的权重共享机制极大地减少了网络模型的参数量, 并以此提升了网络模型的训练效率和泛化能力。然而, 当需要捕捉空间特征的全局关系来识别更复杂的数据模式时, CNN 需要层叠大量的卷积操作来逐步扩大感受野, 这将导致网络模型的参数量急剧上升。

为了解决 RNN 在处理长时间序列时的梯度消失问题以及克服 CNN 在捕捉全局空间特征时的局限性, 研究者们引入了基于自注意力 (Self-Attention) 机制的 Transformer 模型^[64]。与 RNN 不同, Transformer 模型不依赖于顺序处理时间序列, 而是通过自注意力机制同时获取序列中所有位置彼此间的相关性, 这使得 Transformer 模型能够有效捕捉长距离依赖关系, 并且能够避免梯度消失的问题。同时, Transformer 模型相比 CNN 不依赖于局部感受野操作^[65], 而是通过自注意力机制在全局范围内建模空间特征之间的相关性, 这使得 Transformer 模型能够用更少的网络层实现对复杂全局模式的捕捉。然而, 自注意力机制本身并不嵌入先验知识, 而是依赖大量数据来学习有用的表征信息。因此, Transformer 模型需要事先在大量数据上进行预训练 (Pre-Training) 来提升下游任务中的泛化能力。

基于自注意力机制的 Transformer 类模型^[64-67]不仅解决了自然语言处理任务中 RNN 类

模型的顺序依赖性问题,还解决了计算机视觉任务中 CNN 类模型的感受野限制问题,并建立了不同模态数据任务之间的联系。目前,算法领域中的大多数研究工作几乎都是在 Transformer 架构的基础上展开的。

此外,还有一种独特的神经网络——脉冲神经网络^[68] (Spiking Neural Network, SNN)。与上述传统神经网络不同, SNN 的工作机制更接近生物神经网络的工作原理,因此也被认为是第三代神经网络。 SNN 不是像传统神经网络那样通过连续值来传递信息的,而是通过离散的脉冲 (Spikes) 实现信息的传递。 SNN 中的神经元在积累到一定电位时,会发放一个脉冲,类似于生物神经元的“全或无”行为。遗憾的是,由于脉冲神经元的离散性和非线性特性, SNN 的训练比传统神经网络更加复杂,基于 BP 的梯度下降法不能直接应用于 SNN 进行突触权重的更新。为此,研究者们定义了基于脉冲时序依赖可塑性^[43] (Spike-Timing Dependent Plasticity, STDP) 的突触权重更新法则: 1) 时间先后的关系决定突触权重的变化,突触前神经元先发放脉冲则突触权重增加,突触后神经元先发放脉冲则突触权重减小; 2) 时间差的大小影响突触权重变化的程度,当突触前神经元和突触后神经元的脉冲发放时间非常接近时,突触权重的变化幅度会增大,反之,突触权重的变化幅度会减小。然而,这套法则很难适配目前的电子计算机设备,因为生物系统是通过高度并行的方式实现 STDP 法则的,而目前的电子计算机架构^[6,7] 主要是串行或有限并行的。

2.2 人工神经网络的光学实现方式

在类脑光子芯片的硬件层面上直接构建光学神经网络的技术方案多种多样,每种方案都有其独特的结构设计和工作机制。本节将介绍几

种代表性的光学神经网络硬件实现技术^[69-75], 包括 MLP、SNN、CNN、RNN 和衍射神经网络 (Diffractive Neural Network, DNN) 的实现方式 (表 3)。这些技术方案不仅展示了光子芯片在智能化方面的优势,还为未来类脑光子芯片的大规模集成和应用奠定了基础。

1) 光学 MLP。在构建光学 MLP 方面,已经成功应用的技术方案是使用由马赫 - 曾德尔干涉仪组成的阵列来模拟 MLP 的基本网络层,实现矩阵乘法和非线性映射。马赫 - 曾德尔干涉仪方案的代表工作是 Shen 等^[74] 设计的具有 56 个可编程马赫 - 曾德尔干涉仪的相干纳米光子电路 (图 3)。在这项工作中,神经元的物理形态由带热光移相器的光波导组成的马赫 - 曾德尔干涉仪所构成。其中,突触权重矩阵通过奇异值分解 (Singular Value Decomposition, SVD) 的方式被分解为了两个酉矩阵和一个对角矩阵的乘积,这相当于将突触权重矩阵对应的线性变换分解为两次酉变换和一次缩放操作。相比于直接建模复杂的线性变换,酉变换在马赫 - 曾德尔干涉仪系统中可以通过干涉过程描述,而缩放操作可以通过定向耦合器调节光信号的强度来实现。

为了进一步实现神经网络中非线性激活函数

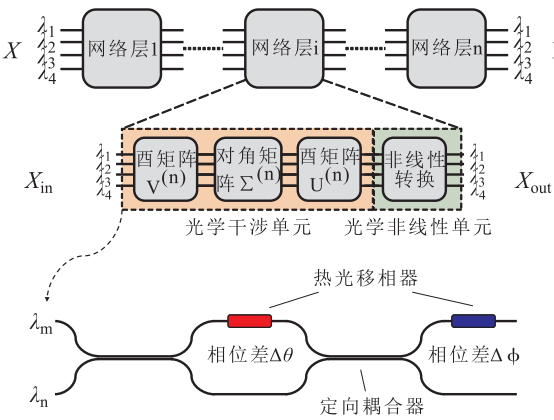


图 3 相干纳米光子电路^[74]

Fig. 3 Coherent Nanophotonic Circuits^[74]

表 3 神经网络的光学实现

Tab. 3 Optical Implementation of Artificial Neural Networks

网络名称	神经元的构造方式	非线性激活函数的实现方式	突触权重更新法则
光学 MLP	带热光移相器的光波导组成的马赫-曾德尔干涉仪 ^[74]	通过在光波导上集成饱和吸收体的方式非线性地调制光	通过电子计算机模拟的方式建模整个系统后执行基于有限差分的梯度下降法调整马赫-曾德尔干涉仪上移相器的相位设置
光 SNN	集成相变材料的光波导和微环谐振器的组合 ^[72,75]	在微环谐振器上集成相变材料使其谐振条件随相变材料相变而发生非线性变化 ^[72]	通过基于 Hebbian 学习规则的反馈机制来改变相变材料的相变程度
光学 CNN	集成相变材料的光波导 ^[75]	网络只在光域上实现卷积计算,非线性映射需要借助电子计算机在电域上进行 ^[71]	通过电子计算机模拟的方式建模整个系统后执行基于 BP 的梯度下降法调整相变材料的相变程度
光学 RNN	基于多节点 基于光电效应的忆阻器 ^[70]	忆阻器基底材料的铁电性和光电特性使得忆阻器的阻值产生非线性变化	输入层到储备池层间的连接权重是随机产生且固定的稀疏矩阵;储备池内部各节点间的连接权重由忆阻器阵列的阻值共同决定;储备池层到输出层间的连接权重通过在电子计算机上应用岭回归算法获取
	基于单节点 半导体激光器通过光学反馈回路生成的虚拟节点 ^[69]	通过引入外部反馈的方式使得半导体激光器产生非线性动力学行为	输入层到储备池层间的连接权重由掩码矩阵决定;各个储备池层之间的连接权重由半导体激光器之间的注入比率和失谐频率决定;储备池层到输出层间的连接权重通过在电子计算机上应用岭回归算法获取
DNN	衍射层中的衍射单元 ^[73]	通过在衍射层上集成光学非线性介质或饱和吸收体的方式非线性地调制光	通过电子计算机模拟的方式建模整个系统后执行基于 BP 的梯度下降法调整衍射单元的几何参数

的功能,Shen 等利用集成石墨烯层的光波导构造了光学非线性单元 (Optical Nonlinearity Unit, ONU)。在光学非线性单元中,作为饱和吸收体的石墨烯在低光强下对光的吸收是线性的。而当入射光强度超过一定阈值时,石墨烯会因泡利阻塞 (Pauli Blocking) 而达到饱和状态。此时,石墨烯对光的吸收率会开始随光强增大而非线性降低。因此,集成石墨烯层的光波导能够根据光强变化实现相应的调制功能。

Shen 等构造的光学 MLP 是典型的光电混合计算架构。其中,记录突触权重的热光移相器依赖于电信号的控制,因此需要借助电子计算机的模拟来实现突触可塑性。虽然尚未实现完全的全光计算,但 Shen 等的方案在片上集成光学神经网络领域具有里程碑式的意义。该方案提出的基于相干纳米光子电路的深度学习架构,实现了

比传统电子系统快两个数量级的前向传播速度,并显著降低了能耗。具体而言,该系统通过并行的光学矩阵乘法来完成大部分计算任务,实现了 100 GHz 的数据处理频率,而传统电子计算机 CPU 的数据处理频率仅为 5 GHz 左右,在相同的能耗下,该系统能比传统电子计算机执行更多的操作。并且,该网络在大规模扩展时依旧能保持低能耗,这在传统电子计算中是难以实现的。这些突破为实现更高效、更低功耗的全光神经网络计算提供了重要的技术基础。Bogaerts 等^[76]在 Shen 等工作的基础上重点分析了马赫-曾德尔干涉仪作为 2×2 模拟光学门的工作原理,提出马赫-曾德尔干涉仪是构建大规模可编程光子电路的核心元件。Harris 等^[77]基于 Shen 等工作探讨了马赫-曾德尔干涉仪实现通用线性光学变换的方式,认为马赫-曾德尔干涉仪为实现复杂的

光学操作提供了一个高效且模块化的实现途径,基于马赫-曾德尔干涉仪的光子神经网络可以用于量子信息处理、机器学习和光通信等领域。

在 Shen 等的方案中存在的主要问题是马赫-曾德尔干涉仪的热光移相器之间存在热串扰。热串扰会影响相位调制的精度,进而影响整个网络的输出精度。针对这类问题,可行的解决途径是使用非易失性的相变材料(Phase-Change Materials, PCM)来构造光学移相器。相比需要持续加热来维持波导温度的热光移相器,基于相变材料的光学移相器只在相变时需要施加电脉冲,无需持续供电驱动,能够在一定程度上规避热串扰的问题。Wei 等^[78]利用相变材料 Sb_2Se_3 和 $\text{Ge}_2\text{Sb}_2\text{Se}_4\text{Te}_1$ 分别设计了两种光学移相器用于实现光子电路中的相位调制。其中,基于 Sb_2Se_3 的移相器具有多级可控的相态(Phase),能够通过精确控制电脉冲的强度或持续时间来实现连续相位调制;而基于 $\text{Ge}_2\text{Sb}_2\text{Se}_4\text{Te}_1$ 的移相器具有较高的消光比,不仅能够实现相位调制,还可调节光强。除了 $\text{Ge}_2\text{Sb}_2\text{Se}_4\text{Te}_1$, Xu 等^[79]研究发现,相变材料 $\text{Ge}_2\text{Sb}_2\text{Te}_5$ 同样可以用于构建高消光比的光学移相器。

2) 光 SNN。2.1 节论述了 SNN 的基本概念,与在电子计算机中不同,以光为信息载体的光子芯片可以利用波分复用的方式实现并行处理,是 SNN 的潜在部署平台。

截至 2024 年 10 月,构建光 SNN 方面的代表工作是 Feldmann 等^[72]提出的基于光子突触与微环谐振器的全光 SNN。对于光子突触的研究,最早是 Cheng 等^[75]将相变材料 $\text{Ge}_2\text{Sb}_2\text{Te}_5$ 通过溅射沉积的方式集成到氮化硅波导上实现的。由于 $\text{Ge}_2\text{Sb}_2\text{Te}_5$ 的相变进程对光脉冲导致的光热效应敏感, $\text{Ge}_2\text{Sb}_2\text{Te}_5$ 可以根据施加的光脉冲数量和强

度在介于完全晶态和完全非晶态之间的多种相态间切换,这些相态具有不同的折射率和吸收特性且是非易失的。

在 $\text{Ge}_2\text{Sb}_2\text{Te}_5$ 的相态中,完全晶态的 $\text{Ge}_2\text{Sb}_2\text{Te}_5$ 具有最高的折射率和吸收率,使得光在传播时受到最大程度的折射和吸收。完全晶态的 $\text{Ge}_2\text{Sb}_2\text{Te}_5$ 可以通过高温熔融-快速淬火的方式向完全非晶态转变, $\text{Ge}_2\text{Sb}_2\text{Te}_5$ 的折射率和吸收率也会逐渐下降,使得光在传播时损耗变小。同时,相变的进程是可逆的,完全非晶态的 $\text{Ge}_2\text{Sb}_2\text{Te}_5$ 可以通过热退火的方式重新转变为完全晶态。这种可逆的相变机制使得 $\text{Ge}_2\text{Sb}_2\text{Te}_5$ 能够将输入光脉冲的性质记录为相变程度,并同时 对光的传播产生影响。因此,基于相变材料 $\text{Ge}_2\text{Sb}_2\text{Te}_5$ 的光子突触能够模拟生物突触的存算一体性。

如图 4 所示,在 Feldmann 等^[72]设计的全光 SNN 中,不同波长的输入脉冲通过光波导传输到特定的光子突触单元中,突触单元依据相变材料的相变程度对这些脉冲进行加权。加权后的输入脉冲经过由微环谐振器和波导组成的多路复用器合并后,通过波导继续传输到一个溅射沉积相变材料的微环谐振器中:这个微环谐振器的谐振条件会随相变材料相变而发生非线性的改变,当材料的相态从非晶态转变为晶态时,折射率增大,进而导致谐振波长向长波长方向偏移;反之,当材料由晶态转变为非晶态时,折射率减小,谐振波长向短波长方向偏移。谐振波长发生偏移后,如果谐振波长与输入光的波长更加接近,则微环谐振器 drop 端口处的传输增大;如果谐振波长远离输入光的波长,则微环谐振器 drop 端口处的传输减小。因此,将相变材料与微环谐振器相结合能够实现对光信号的加权与非线性调制。

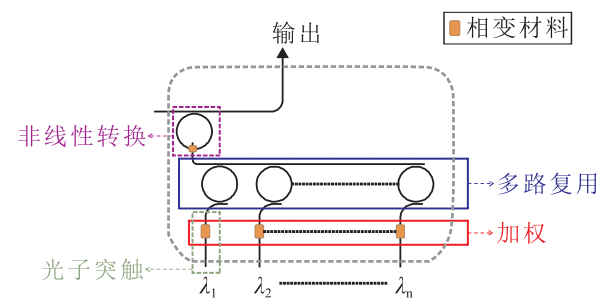


图4 基于光子突触的全光 SNN 的一个基础网络层^[72]

Fig. 4 A Basic Layer in All-optical Spiking Neurosynaptic Network^[72]

为了训练这样一个全光 SNN, Feldmann 等设计了一种反馈机制:输出层输出的脉冲信号会通过反馈波导作为反馈脉冲返回到输入层的光子突触单元中。当反馈脉冲与特定波长的输入脉冲发生重叠时,脉冲能量增大,对应光子突触单元上相变材料的相变程度因为光热效应发生改变,使得该波长的输入脉冲权重变大。这种权重更新机制遵循了 Hebbian 学习规则^[42],即共同激发的神经元会加强连接。

Feldmann 等设计的方案不仅通过波分复用技术实现了高度并行的数据处理流程,而且通过不依赖电信号的非易失性相变材料实现了存算一体架构,是光学系统实现类脑计算的重大突破。其中,相变材料 $\text{Ge}_2\text{Sb}_2\text{Te}_5$ 是该方案中实现光 SNN 的核心。然而,这类相变材料的耐久度不高,仅能够实现 10^{12} 次相态切换,在长时间使用后可能导致网络性能下降。并且,转换相变材料相态的能耗很高,每次材料的晶化或非晶化过程需要约 660 pJ 的能量。相态转换速度也比较慢,在使用较低能量的光脉冲修改光子突触权重时需要 65 个操作周期才能将权重重置零。后续工作应当在相变材料的选取上寻求进一步的突破,使得切换相态所需的能量尽可能少,进而使光子突触在突触可塑性上实现与人脑相近的能耗。

与集成相变材料不同,也有一些研究工作将热光效应引入到微环谐振器系统中。Tait 等^[80]通过施加电流来加热微环结构,使其温度升高。在这过程中,硅的折射率会随着温度的升高而非线性地增加,使得微环的谐振波长向长波长方向偏移,从而实现加权和非线性调制。同时,Tait 等^[81]也实现了波分复用。然而,这种由热光效应产生的加权与非线性调制,不可避免地会引入热串扰,从而影响加权与非线性调制的精度。

3) 光学 CNN。近年来,基于相变材料的光子突触被用于构建光学 CNN。在 CNN 中,卷积层中的卷积操作是通过卷积核 (Convolution Kernel) 在输入数据的不同数据范围内依次执行一个点积运算来完成的。这个点积运算可以被表示为一个矩阵 - 向量乘积 (Matrix-Vector Multiplication, MVM), 卷积操作实际上就是多个矩阵 - 向量乘积的组合。对此,Feldmann 等^[71]定义了一种每秒能执行数万亿次乘积累加运算 (Multiply Accumulate, MAC) 的集成光子张量核心 (Integrated Photonic Tensor Core), 用于并行化地计算矩阵 - 向量乘积,从而在硬件层面加速 CNN 中的卷积操作。通过该项技术,每平方毫米网络的计算密度可达到每秒 400 万亿次运算,并且网络每次乘积累加运算的功耗最低可达到 17 fJ,这使得网络在能效上极具优势。

在集成光子张量核心中,卷积核被定义为了物理层面上由光子突触组成的阵列。其中,卷积核的参数被编码为了光子突触上相变材料的相变程度。如图 5 所示,利用多波长相干光源和光子突触阵列实现波分复用是集成光子张量核心进行高度并行化计算的关键。

Feldmann 等^[82]选用了光频率梳 (Optical Frequency Comb) 作为多波长相干光源。光频率

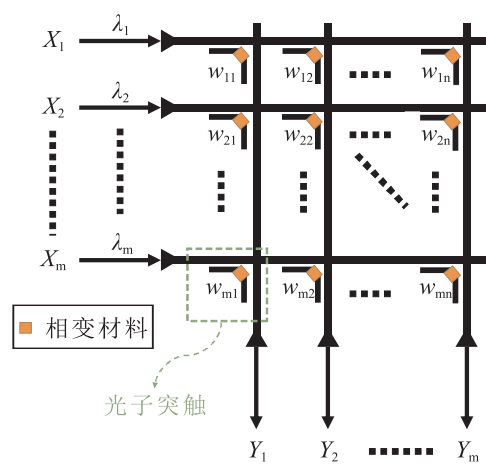


图 5 基于光子突触的集成光子张量核心^[71]

Fig. 5 Integrated Photonic Tensor Core Based on Photonic Synapses^[71]

梳是由一系列等间隔的光谱线组成的光谱,其中每条谱线都对应一个特定的波长。与使用多个连续波激光器(Continuous-Wave Laser)作为光源时所面临的校准问题相比,光频率梳不仅波长间隔稳定,而且可以通过单一的连续激光泵浦高品质因子的微环谐振器产生,这在很大程度上降低了多波长相干光源系统制备的复杂性,有利于片上的集成。

遗憾的是,该工作仅用光学方法实现了 CNN 中卷积层的功能,非线性映射、突触可塑性以及 CNN 中其他网络层的功能都需要借助电子计算机的模拟来完成。尽管 Feldmann 等的工作未能通过光学方法实现 CNN 中所有网络层的功能,但该工作是高效并行处理与非易失性光学存储技术相结合思想在硬件加速方面的重要应用,为未来数据密集型计算任务中光子芯片的设计思路提供了重要参考。

受到 Feldmann 等工作的启发,集成光频率梳技术有望成为光子芯片的全新设计范式。Bai 等^[83]在将光频率梳应用于光学神经网络后发现,

光频率梳本身具有的多个等间隔频率作为参考信号,不仅降低了系统对调制速度的要求,还使系统能够处理更大的数据负载并获得更好的计算性能。Shu 等^[84]使用光频率梳驱动硅光子系统,解决了传统硅光子电路缺乏稳定多波长光源的局限性。研究发现,光频率梳技术的应用降低了光学元件的片上集成难度,有助于将系统缩小至芯片级。可见,光频率梳是片上系统集成和扩展的核心组件,能够突破光计算的瓶颈。

4) 光学 RNN。储备池计算 (Reservoir Computing, RC) 网络是一种利用具有随机连接权重的非线性动力系统 (Dynamical System) 来构建隐藏层 (储备池) 的 RNN。与传统 RNN 相比,储备池计算网络中只有隐藏层到输出层间的连接权重需要通过训练进行学习,这大大简化了计算过程,规避了传统 RNN 因训练循环连接权重而带来的高计算开销问题,十分利于 RNN 的硬件化。

基于多节点系统的储备池计算 (图 6a) 是通过互连大量具有非线性动态响应的物理节点来构建储备池的。目前,这类储备池计算网络主要由忆阻器 (Memristor) 组成的阵列所构成,因为忆阻器是一种具有电荷记忆功能的非线性电阻,能够作为储备池中的物理节点提供储备池计算所需的短期记忆功能和非线性映射。

多节点系统在光学方面的代表工作是 Liu 等^[70]提出的基于光电忆阻器的储备池计算网络。在该方案中,光电忆阻器作为突触结构利用了 α - In_2Se_3 材料的光电性质实现了突触可塑性:当光照射到以 α - In_2Se_3 为基底的光电忆阻器时,光生电流会迅速增加,然后随着时间的推移逐渐衰减,这种电流响应与生物突触在接收到神经递质后产生的突触后电流 (Postsynaptic Current, PSC) 的响应形式非常相似,模拟了短时程增强 (Short-

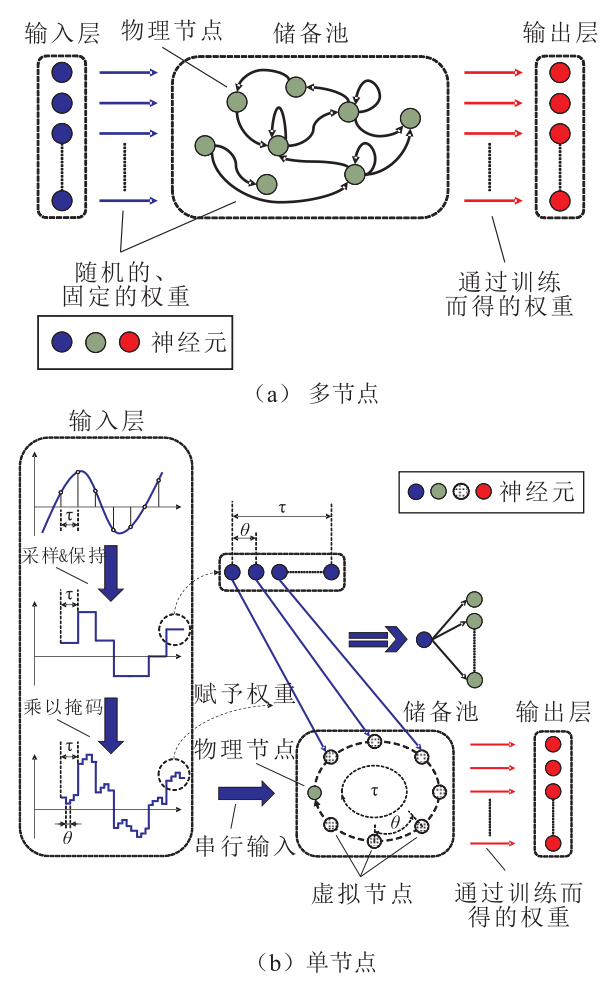


图 6 储备池计算的实现方案^[85]

Fig. 6 Reservoir Computing Schemes^[85]

term Potentiation, STP)。同时, Liu 等施加了一对时间间隔较短的光脉冲到光电忆阻器上。实验表明,当第二个光脉冲到达光电忆阻器时,光生电流会相对于第一个脉冲到达时的响应有所增强,这表明这种光电忆阻器能够模拟突触可塑性中的双脉冲易化(Paired-pulse Facilitation, PPF)现象。

除了光电性质, $\alpha\text{-In}_2\text{Se}_3$ 材料还具有独特的铁电性(Ferroelectricity),能够在施加电场后产生并保持自发极化。这种极化的方向和程度可以通过外加电场来调节,从而影响材料的电导状态。因此,光电忆阻器可利用铁电性进一步完善

突触可塑性:当重复施加电压脉冲时,材料中的铁电极化逐渐积累并导致器件的导电通道无法自发恢复至初始状态,从而实现类似于生物突触的长时程增强(Long-term Potentiation, LTP)效应。同时,相比光脉冲,电压脉冲还存在正负之分,因此光电忆阻器还能实现短时程抑制(Short-term Depression, STD)和长时程抑制(Long-term Depression, LTD)。

结合 $\alpha\text{-In}_2\text{Se}_3$ 材料的光电特性和铁电性,基于 $\alpha\text{-In}_2\text{Se}_3$ 的光电忆阻器能够同时处理光信号和电信号,并且通过调节背栅电压(Back-Gate Voltage)还可以改变光电忆阻器的弛豫时间(Relaxation Time),进而使得光电忆阻器对不同变化速率的信号敏感。因此,基于光电忆阻器的储备池计算网络能够完成多模态和多时间尺度的计算任务。

虽然全光计算备受关注,但 Liu 等的研究充分展示了光电混合架构在类脑计算中的潜力,尤其是在多模态和多尺度信息处理方面的应用, Liu 等设计的储备池计算网络在多模态手写数字识别任务中达到了 98.6% 的识别准确率,并且能够在带有 25% 噪声的环境中识别二维码。可见,基于光电忆阻器的光电混合架构结合了光学和电子元件的优势,能够通过光电耦合作用实现高效且鲁棒的信息处理。

将线性不可分的数据映射到更高维度的空间中可以增加数据线性可分的可能性^[61]。在多节点系统中,储备池中的每个节点都有物理实体与之对应,即储备池的状态空间是有限的,这不利于复杂非线性问题和高维数据的处理。对此, Appeltant 等^[85]提出了解决方案:将多节点系统替换为基于时间延迟反馈的单节点系统(图 6b)。

在单节点系统中,储备池仅有一个物理节

点,该物理节点映射的信号会在延迟 τ 之后作为反馈信号耦合到该节点的输入信号上,形成反馈回路。在反馈回路中,间隔 θ 进行一次采样,就可获得反馈回路不同时刻的瞬时状态。这些瞬时状态是对输入信号及其历史信息的复杂非线性映射,因此可视作描述储备池状态的虚拟节点,这相当于在反馈回路中插入了 $N = \frac{\tau}{\theta}$ 个物理节点。

为了构建这种反馈回路,输入信号被采样并保持了一个延迟间隔 τ 。在单个延迟间隔 τ 内,被采样的信号以 θ 为时间步串行输入到同一个物理节点。每当一个时间步上的信号被物理节点非线性映射时,就可看作原输入信号与一个虚拟节点产生了连接。其中,单个延迟间隔 τ 内的输入信号到储备池中各个虚拟节点的连接权重是通过乘以掩码的方式实现的。具体来说,掩码会改变输入信号在延迟间隔 τ 内的不同时间步 θ 上的取值,使得每个时间步上的信号在输入到同一个物理节点进行非线性映射前就乘以了相应的权重。

截至 2024 年 10 月,单节点系统在光学方面的最新进展是 Shen 等^[69]设计的一种基于半导体

激光器的深度光子储备池计算网络。与其他方案^[86,87]相比,Shen 等的方案首次在不借助光电转换和模数转换的情况下使用全光的方式实现了多个储备池层的级联。

如图 7 所示,深度光子储备池计算网络由一个主激光器 (Master Laser) 和四个储备池通过级联的方式所构成。其中,每个储备池是由一个从激光器 (Slave Laser) 和一个光学反馈回路所构成的循环连接结构。在深度光子储备池计算网络中,主激光器的输出光会通过光注入 (Optical Injection) 的方式逐层传递到从激光器,每一层的激光器都会被锁定在与上一层激光器相同的频率和相位上。这种级联的光注入锁定确保了层与层之间信号的同步性,使得每一个储备池层都可以在光域中直接接收和处理光信号,而不需要将光信号转换为电信号,是多个储备池层之间实现全光级联的基础。同时,前后两层激光器之间的注入比率以及失谐频率决定了储备池层之间的互连权重。

在 Shen 等构建的单节点系统中,虚拟节点个数以及掩码矩阵是由任意波形发生器 (Arbitrary Waveform Generator, AWG) 和光调制器的组合实

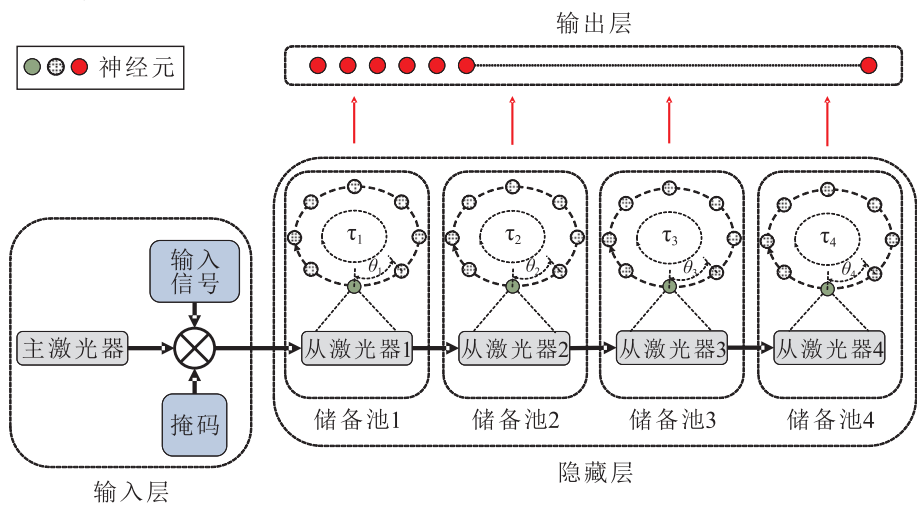


图 7 深度光子储备池计算网络^[69]

Fig. 7 Deep Photonic Reservoir Computing^[69]

现的。每个储备池中的半导体激光器作为物理节点,发出的光信号经过耦合器被分为两部分,一部分继续向下一层输出,另一部分通过光环行器(Optical Circulator)作为反馈信号被引导回激光器。通过引入光反馈信号,半导体激光器会产生不稳定的输出,例如单周期振荡、倍周期振荡和混沌等非线性动态特性,从而生成复杂的非线性映射,实现网络中激活函数的功能。

然而,在单节点系统中,虚拟节点的输出是按时间顺序串行生成的,这意味着每个节点必须在前一个节点输出完成后才能处理信号。而在多节点系统中,多个物理节点间没有时间依赖性,可以并行地处理输入信号。因此,相比多节点系统,单节点系统虽然能够利用单一的物理节点生成大量虚拟节点来降低系统的硬件复杂性和成本,但也存在因串行处理而导致的速度限制。

5)DNN。DNN 是一种利用光的衍射效应来模拟传统神经网络计算过程的光学计算系统。如图 8 所示,DNN 是由一系列依次排列的衍射层所构成的。其中,衍射层是由许多离散的衍射单元(比如一个小孔或一个微小的结构)所组成的平面,从衍射层平面透射的相干光会发生衍射。此时,衍射层中的每个衍射单元可以看作是一个能够发出次波(Secondary Wave)的波源。并且,次波波前(Wavefront)上的面元分布描述了前后衍射层中衍射单元间的指向性连接点。因此,DNN 可以模拟 MLP 的全连接结构。其中,衍射层中的每个衍射单元都与上一层中的所有衍射单元以波的形式全连接。

当光波在两个衍射层间进行传播时,前一个衍射层上的微小结构会改变光波相位或振幅的分布,使得光波在后一个衍射层即接收面上形成特定的光场分布。通过调整前一个衍射层上的

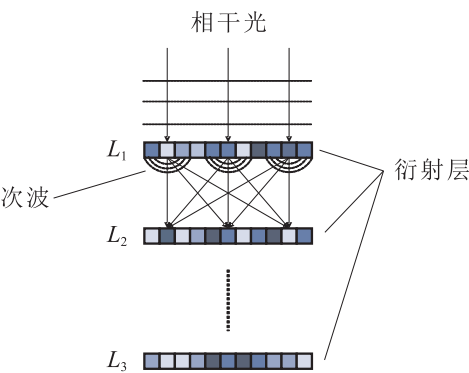


图 8 DNN 示意图^[73]

Fig. 8 Illustration of Diffractive Neural Network^[73]

微小结构,就可以控制光场的分布。与神经网络的网络层类似,衍射层可以看作是一个矩阵,其中的矩阵元素代表衍射层中衍射单元的传输函数,即对光波相位或振幅的调制。衍射层矩阵描述了光场通过衍射层后的线性变换关系,衍射层输出的光场可以看作是输入光场与衍射层矩阵进行矩阵乘法后得到的结果。

DNN 方面的代表工作是 Lin 等^[73]通过 3D 打印制作的衍射深度神经网络(Diffractive deep neural networks,D²NN)。在 Lin 等的方案中,衍射层的衍射单元是一种基于 3D 打印材料 VeroBlackPlus RGD875 的微米级结构。这些微米级结构以阵列的方式排布在 3D 打印材料的基板上,当光波通过具有不同厚度的材料时,光波的波前会发生相位延迟,从而改变其传播方向,实现对输入光场的调制,并最终在网络的输出平面生成所需的光强分布来完成特定的分类或成像任务。在这过程中,每个微米级结构的厚度作为网络的权重值是由电子计算机模拟并建模整个系统后通过执行 BP 算法获取的。同时,计算机模拟建模过程中使用的非线性激活函数可以通过在衍射层上集成光学非线性介质或饱和吸收体的方式实现。

Lin 等的研究通过引入 D^2NN 架构,提出了一种全光机器学习的设计方案。该方案规避了复杂光波导的设计,极具成本效益。具体来说, D^2NN 网络中的神经元是无源的通过 3D 打印制成的微结构,整个网络的扩展仅需增加额外的 3D 打印层。这意味着可以通过简单的制造工艺实现大规模的神经网络部署,而无需复杂的光路设计。这种经济高效的可扩展性使得 D^2NN 能够适用于各种复杂的计算任务,例如图像分类、特征提取和目标检测等应用场景。通过整合高通量、大面积的 3D 打印层, D^2NN 有望达到数亿神经元和数千亿连接的网络规模,从而以低成本和高能效的方式推动全光计算的发展。

然而,在 D^2NN 架构中,衍射层之间的相对位置和对齐精度会直接影响网络的性能。如果衍射层之间的对齐出现微小偏差,光在层间传播时的衍射路径就会偏离设计预期,导致输出结果的不准确。因此,尽管 D^2NN 在理论上具备高度扩展性,但在实际应用中,需要精确的对齐控制以确保网络能够保持较高的计算精度。这种对齐要求增加了网络设计的复杂性,是限制 DNN 发展的一个主要挑战。

针对 DNN 衍射层的对齐问题,Wu 等^[88]提出了一种只训练输入层中的电-光转换过程,并使用单一的衍射层进行图像分类的网络架构。整个网络只使用一个衍射层,这大大简化了光学系统的复杂性,避免了衍射层之间因对齐问题而产生的误差。然而,虽然该架构通过仅训练输入层的电-光转换简化了系统,但这也意味着其整体计算能力较为有限。相比于多层的 DNN,它无法利用更多的网络层次来进行复杂的计算,这会限制其在处理复杂任务时的能力。Lou 等^[89]探究了层间反射和像素间相互作用对 DNN 性能的影响

后指出,层间反射对网络的分类准确率有显著影响。高折射率材料会导致较大的光反射,从而增加系统误差;像素间的相互作用会导致解析模型的计算结果和实验的实际结果间存在偏差。为了解决这些问题,该项工作引入层间反射的计算,改进了对真实硬件系统的模拟,并采用时域有限差分法 (Finite-Difference Time-Domain, FDTD) 来替代传统的解析模型。通过结合上述两种方法,网络能够预测实验中由于反射引起的功率损失,并降低像素间的相互干扰。

3 类脑光子芯片的发展瓶颈

3.1 非线性光学元件的匮乏

大脑是一个高度复杂的系统,其中涉及大量的非线性动力学过程^[90]。为了模拟脑的非线性特性,人工神经网络引入了隐藏层和非线性激活函数。正如 2.1 节中论述的那样,如果神经网络中每一层都是线性的,那么无论网络层数有多少,整个网络都可以简化为一个线性变换。这意味着网络的输出是输入的线性组合,无法在高维空间中将线性不可分的输入映射为线性可分的输出。因此,引入非线性光学元件是基于模拟光计算的类脑光子芯片能够构建光学神经网络,并具备模拟大脑复杂计算能力的关键。

在光学神经网络中,光信号通过多级处理的方式在网络中传递和转换。网络中每一级的元件都会对光信号进行某种形式的调制,使得整个网络能够通过这些级联的处理步骤完成复杂的计算任务。为了实现这种级联性,网络中某一级元件的输出信号强度必须足以驱动下一级的元件,否则信号会逐级衰减,影响整个系统的性能和可靠性。然而,光子之间的直接相互作用非常弱,导致在一般条件下光子的非线性响应也相对

较弱,不足以驱动下一级的光学元件。这意味着非线性光学元件需要与大功率激光器相匹配来激发显著的非线性响应并以此驱动下一级元件。大功率激光器造成了较高的能耗,并且高能耗导致的散热问题进一步阻碍了元件的小型化,使得基于光与介质非线性效应的全光非线性操作在需要低功耗和高集成度的芯片应用中面临挑战。

为了规避使用大功率激光器,光-电-光(Optical/Electrical/Optical, O/E/O)转换被广泛用于实现光学神经网络的非线性操作。在O/E/O转换中,首先,光信号通过光电探测器(如光电二极管)转换为电信号,完成光-电转换;其次,转换的电信号可以通过传统的电子电路进行处理来实现各种非线性操作,例如借助二极管的伏安特性、晶体管的开关特性或者非线性放大器等电路实现;最后,非线性处理后的电信号再通过电光转换器(如激光器、调制器等)转换回光信号,完成电-光转换,以便继续在光学神经网络中传播和处理。

O/E/O转换通过混合使用光学和电学元件的方式解决了光学系统中难以直接实现非线性操作的问题。然而,每次光-电和电-光转换都会引入一定程度的噪声,噪声的累积可能对高精度计算和信号处理产生不利影响。并且,转换过程引入了时间延迟,当系统需要高速处理时,O/E/O转换的速度可能成为瓶颈。

因此,无论是全光架构还是光电混合架构,目前在构建类脑光子芯片的过程中仍然缺乏一种能够低功耗激发且响应速度快的非线性光学元件来实现非线性操作。非线性光学元件的匮乏限制了类脑光子芯片的性能,并阻碍了类脑光子芯片的进一步发展。

为了克服光子芯片设计中非线性元件匮乏

的局限性,研究者们发展了多种解决方案。

1)开发 III-V 族半导体材料。III-V 族半导体材料,如砷化镓(GaAs)、磷化铟(InP)等,具有较强的非线性效应,是制作高效非线性光学元件的潜在材料。Kim 等^[91]设计了一种基于绝缘衬底铝镓砷(AlGaAs-on-Insulator, AlGaAsOI)的微环谐振器作为非线性元件,通过 Kerr 效应实现非线性调制。其中,AlGaAsOI 的非线性折射率系数远大于氮化硅 Si₃N₄ 和氮化铝 AlN 等常见波导材料,能够在低功率下实现强非线性效应。然而,AlGaAsOI 这类具有高折射率对比度的波导系统更容易受到表面粗糙度散射损耗的影响,因此需要高度精确的制造工艺,如电子束光刻^[92]和干法刻蚀^[93],以确保波导表面光滑。

2)开发铌酸锂材料。铌酸锂(LiNbO₃)材料具有非常高的二阶非线性系数,频率转换效率高,因而对泵浦源功率的需求较低,十分利于片上集成实现非线性调制。然而,类似于 AlGaAsOI,基于铌酸锂的波导系统也具有高折射率对比度,因此其早期在集成光学中的发展受到波导制备工艺的阻碍^[94]。随着薄膜铌酸锂^[95](Thin-Film Lithium Niobate, TFLN)技术的突破,铌酸锂可被制备成亚微米级的薄膜,这极大地推动了集成光学的发展。目前,基于智能切片^[96](Smart Cut)工艺制备而成的薄膜铌酸锂被广泛用于构建集成光学元件。在利用薄膜铌酸锂构建非线性元件时,相位匹配条件是影响频率转换效率的关键因素。在频率转换过程中,需要确保参与非线性相互作用的光波在波导中保持相位同步。如果没有良好的相位匹配,频率转换效率就会急剧下降。然而,光波在普通的铌酸锂材料中天然地达成相位匹配通常很难实现,或只能在有限的波长范围内有效^[97]。对此,研究者们通过

设计周期性极化反转结构进行准相位匹配^[98],使得相位失配被周期性补偿,从而增强非线性效应,这正是 1.1.1 小节中提及的周期性极化铌酸锂。

3) 开发钙钛矿材料。卤化物钙钛矿 (Halide Perovskite)^[99] 因其优异的非线性光学特性而受到广泛关注,这类材料具有较高的非线性光学系数、较宽的带隙可调性和较强的光学吸收能力,是构建非线性光学元件的理想选择。例如, Walters 等^[100] 利用甲基铵铅溴化物 ($\text{CH}_3\text{NH}_3\text{PbBr}_3$) 钙钛矿单晶的双光子吸收特性构建了非线性光学元件。卤化物钙钛矿材料可以通过溶液法在低温下制备,生产工艺相对简单且成本较低。相比传统半导体材料,钙钛矿的制备不需要复杂的设备和高温条件,非常适合大规模生产。然而,钙钛矿材料在实际应用中仍面临稳定性问题,其对湿度和温度相对敏感。为了克服这些挑战,研究者们正在探索通过材料工程和封装技术来提高其环境稳定性^[101],以满足光子芯片集成的需求。

4) 开发饱和吸收体。饱和吸收体的吸收特性随着入射光强度的增加而非线性降低,这一特性使其在被动调 Q 和锁模激光器中得到了广泛应用^[102]。目前,在使用饱和吸收体构建光学非线性元件方面,主流的做法是将石墨烯作为饱和吸收体,集成到波导或者微环谐振器上^[103],实现非线性调制。然而,石墨烯作为饱和吸收体用于高功率非线性光学应用时,存在功率限制的问题。研究表明^[104],在高光功率下,石墨烯涂层元件的非线性特性会发生不可逆的退化,这限制了其在高功率环境中的应用。

5) 开发相变材料。相变材料是一类能够在不同相态之间可逆转换的材料,其光学性质在相

变过程中会发生显著变化,十分适合构建非线性光学元件。与饱和吸收体类似,将相变材料集成到光波导结构上后,可以通过控制相变材料的相态来调制波导的有效折射率,从而实现非线性调制。然而,相变过程需要一定的能量和时间^[105],这限制了器件的功耗和响应速度,并且相变材料在多次循环相变后可能出现性能退化^[106]。

6) 混合集成和异质集成。近年来,非线性光学领域的一个重要研究方向是将石墨烯等二维材料与硅和氮化硅等集成平台通过混合集成的方式相结合^[107],开发出更高效的片上非线性光学元件。除了混合集成,在一个单一系统中,将不同材料构成的非线性元件进行异质集成,也能够增强非线性系统的整体性能。例如,铌酸锂薄膜与硅氮化物波导层的异质集成展示了在片上实现高光学非线性和低损耗的潜力^[108]。

3.2 网络规模扩展的挑战

在电子计算机的软件层面上,基于深度学习技术的人工神经网络模型通过多层网络结构和大量的权重参数来处理复杂的任务。网络层数越深,模型能够学习到的数据表示就越复杂和抽象。因此,当处理图像识别、自然语言处理和语音识别等复杂任务时,神经网络模型通常需要较多的网络结构来逐步提取数据的高阶特征。同时,这种深层网络模型依赖大量的权重参数来调整网络层之间神经元的连接,使得模型能够更好地拟合复杂的模式和数据分布。随着网络层数和参数规模的增加,深度神经网络模型具备了更强的学习能力,但也带来了更高的计算开销。

尽管类脑光子芯片实现的光学神经网络在能效和处理速度方面具有明显优势,但它们在处理复杂任务时面临着网络规模扩展的限制,无法进一步提升智能化水平。由于模拟系统对噪声

敏感,简单地通过扩展类脑光子芯片中光学神经网络的网络层数来提升芯片的智能化水平是不切实际的,因为模拟噪声会随着网络层数的增加而指数级增长,从而导致芯片性能下降。基于这个原因,2.2 节中讨论的大部分光学神经网络是浅层模型,仅能处理较为简单的任务,并且部分网络是不可重配置的^[73],即网络一旦训练完毕就不可再修改权重,仅适用于固定的任务。

2024 年 4 月,清华大学戴琼海院士团队设计了一种基于分布式计算架构的太极芯片^[109],实现了千万类别级的大规模图像分类任务以及高保真的多模态内容生成任务,是类脑光子芯片构建的光学神经网络突破网络规模瓶颈的里程碑。在太极芯片构建的光学神经网络中,复杂的任务被分布式处理策略拆分为了多个简单的子任务,并且这些子任务被分配给多个独立的处理单元并行处理。这种独特的分布式计算架构减少了网络误差的积累,使得光学神经网络规模的扩展成为了可能。并且,多个子任务并行处理的方式进一步提高了芯片的处理速度,使得太极芯片超越了同算力级别的电子芯片架构。

为进一步增强芯片的可重配置性,戴琼海院士团队开发了一种摒弃 BP 的全前向模式(Fully Forward Mode, FFM)学习准则^[110],使得太极芯片不再依赖电子计算机的模拟进行训练。FFM 学习准则利用空间对称性和洛伦兹互易性,以前向传播的方式计算误差相对于网络权重的梯度,实现了对 BP 的等效替代。这种方法通过在物理光学系统中直接进行梯度计算的方式,大大减少了对电子计算机模拟的依赖,避免了离线模型和真实系统之间不匹配而导致的模拟不准确问题,提高了训练的效率和精度。

遗憾的是,太极芯片中非线性激活函数的实

现以及网络权重的更改,依旧需要借助电信号控制,并不是真正意义上的全光类脑光子芯片。并且,太极芯片构建的光学神经网络层数也只达到了 5 层,与电子计算机软件层面能够实现的人工神经网络层数相比,仍有不小的差距。

太极芯片的问世是在硬件层面扩展网络规模取得的重大突破。此外,戴琼海院士团队^[111]还提出了一种在数学模型层面上等效扩展网络规模的方式。该项工作利用常微分方程(Ordinary Differential Equation, ODE)来参数化神经网络隐藏层的连续动态。它的核心理念是通过 ODE 求解器来处理光学神经网络的层间状态变化,并通过时间步长的调控,而不是传统神经网络中堆叠离散层的方式,来模拟多层网络的处理效果,从而在不增加物理层数的情况下,通过反馈回路和积分操作来实现更深的网络效果。具体而言,这项工作引入的光子积分器和光学反馈回路使得网络的输出能够不断反馈到输入进行迭代处理,这相当于在同一物理层上,通过时间上的反复计算来模拟更多层的网络处理效果。

综上所述,类脑光子芯片在网络规模扩展方面取得了重要的进展,尤其是在能效和计算规模上的突破,但芯片在噪声控制、深度架构支持以及大规模硬件实现方面仍存在显著挑战。当前的技术尚不足以支持类脑光子芯片在硬件层面上直接实现电子计算机软件层面定义的深度学习架构。绝大多数类脑光子芯片构建的光学神经网络为浅层网络,而不是深层网络。浅层的网络结构设计虽然能有效减少误差的积累,但无法充分利用深度学习技术的优势。这种网络规模扩展的瓶颈限制了类脑光子芯片在复杂任务上的应用,并阻碍了类脑光子芯片智能化水平的提升。

4 类脑光子芯片研究展望

随着人工智能生成内容 (Artificial Intelligence Generated Content, AIGC) 技术的快速发展, 各类深度学习模型对算力的需求呈现出爆发式增长。这些模型在处理图像、文本和语音等数据时, 往往需要极其庞大的计算资源。然而, 目前的电子芯片制造工艺面临物理极限的挑战, 使得电子芯片在性能上难以满足深度学习模型的需求。为了解决这一困境, 类脑光子芯片的研发成为了关键。发展类脑光子芯片不仅能够突破现有电子芯片的技术瓶颈, 还能为我国实现芯片自主化提供强有力的支持。类脑光子芯片技术难点的突破将为 AIGC 技术的发展带来质的飞跃, 并推动我国科技产业迈向新的高度。

1) 进一步探索非线性光学元件, 构建全光非线性操作层。非线性的激活函数是神经网络能够表示和学习复杂非线性关系的关键因素。在电子计算机的软件层面上, 实现非线性激活函数相对简单且灵活。神经网络中的每个神经元在接受到输入信号后, 会通过数学计算进行非线性映射, 这个过程主要依赖于软件中的数值运算。然而, 在光子芯片的硬件层面上, 实现这种非线性激活函数则困难得多。如 3.1 节中所述, 目前主流的做法是将光信号转换为电信号后再进一步实现非线性处理, 但这一过程存在延迟, 难以满足绝大多数应用场景的高速处理需求。考虑到钙钛矿^[112] 和铌酸锂^[113] 等非线性光响应材料的技术突破, 后续研究应当更关注于实现基于光与介质非线性效应的全光非线性操作。

2) 重点研究大脑以记忆为基础的学习认知原理, 设计匹配光子特性的类脑智能算法。大脑作为自然界中最有效的计算系统, 能够在低能耗

下实现高度并行的计算和信息存储功能。这得益于其在神经元和突触层面上的紧密耦合, 使得计算和存储过程在同一结构中同步进行, 避免了传统计算系统中存算分离所带来的延迟和能耗问题。光子的独特特性展示了它在构建存算一体网络中的巨大潜力: 相比电子, 光子具有高速传输、低功耗和多通道并行处理的优势, 非常适合在硬件层面模拟大脑中神经元和突触的紧密耦合特性。通过结合智能光子技术与生物脑智能研究成果来设计匹配光子特性的类脑智能算法, 类脑光子芯片不仅能够在效率和能耗上更接近生物脑, 还能突破现有深度学习框架下网络规模扩展的限制, 处理复杂数据和模式识别任务。因此, 研究大脑的记忆学习原理并以此构建不同于现有深度学习架构的存算一体光学神经网络, 是未来类脑光子芯片取得突破性进展的核心之一。

参考文献

[1] LECUN Y, BENGIO Y, HINTON G. Deep Learning [J]. Nature, 2015, 521 (7553) : 436-444.

[2] ROMERA-PAREDES Bernardino, BAREKATAIN Mohammadamin, NOVIKOV A, et al. Mathematical Discoveries from Program Search With Large Language Models [J]. Nature, 2024, 625 (7995) : 468-475.

[3] KIRILLOV A, MINTUN E, RAVI N, et al. Segment Anything [C]// ICCV. Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023 : 4015-4026.

[4] ACHIAM J, ADLER S, AGARWAL S, et al. GPT-4 Technical Report [EB/OL]. (2024-03-04) <https://arxiv.org/abs/2303.08774>.

[5] SEFF A, CERA B, CHEN D, et al. Motionlm:

Multi-Agent Motion Forecasting as Language Modeling [C] // ICCV. Proceedings of the IEEE/ CVF International Conference on Computer Vision. Paris: IEEE, 2023: 8579-8590.

[6] ARIKPO II, OGBAN FU, ETENG IE. Von Neumann Architecture and Modern Computers [J]. Global Journal of Mathematical Sciences, 2007, 6(2): 97-103.

[7] LI Bin, LI Wenfeng. Modeling and Simulation of Container Terminal Logistics Systems Using Harvard Architecture and Agent-Based Computing [C] // WSC. Proceedings of the 2010 Winter Simulation Conference. Baltimore: IEEE, 2010: 3396-3410.

[8] HUO Fupeng, CHEN Chuantong, ZHANG Zheng, et al. Development of Thermal Characteristics Evaluation System for Multi-Chip SiC Power Modules [C] // ICEP. 2024 International Conference on Electronics Packaging. Toyama: IEEE, 2024: 47-48.

[9] BHARDWAJ H, JAIN S, SOHAL H. Energy Optimization for RC and RLC Interconnect Design in Low Power VLSI [J]. Micro and Nanosystems, 2024, 16(1): 26-35.

[10] HONG T, KIM S, PARK J, et al. Guidelines for Area Ratio Between Metal Lines and Vias to Improve the Reliability of Interconnect Systems in High-Density Electronic Devices [J]. Electronics, 2023, 12(21): 4403.

[11] FERNANDES C, DE OLIVEIRA M, ROCHA D, et al. Design of Optical Logic Gates Using Mach-zehnder Interferometers and Machine Learning [J]. Journal of Lightwave Technology, 2022, 40(18): 6240-6248.

[12] TIAN Yonghui, ZHANG Lei, JI Ruiqiang, et al. Proof of Concept of Directed OR/NOR and AND/NAND Logic Circuit Consisting of Two Parallel Microring Resonators [J]. Optics Letters, 2011, 36(9): 1650-1652.

[13] QIU Ciyuan, XIAO Huifu, WANG Liheng, et al. Recent Advances in Integrated Optical Directed Logic Operations for High Performance Optical Computing: A Review [J]. Frontiers of Optoelectronics, 2022, 15(1): 1.

[14] PERRON N, CHEMNITZ M, FISCHER B, et al. THz-Bandwidth Passive Logic Gates Through Noninstantaneous Nonlinearity [C] // CLEO. 2023 Conference on Lasers and Electro-Optics Europe & European Quantum Electronics Conference, 26-30 June 2023. Munich: IEEE, 2023: 1-1.

[15] MICHAEL M, CAROLINE B, VIDHYA J, et al. All-Optical Logic Gate Using Photonic Crystals for Ultra-Fast Telecommunication Applications [J]. Photonic Materials: Recent Advances and Emerging Applications, 2023: 21.

[16] ALAM B, CESCHINI A, ROSATO A, et al. All-Optical AND Logic Gate Based on Semiconductor Optical Amplifiers for Implementing Deep Recurrent Neural Networks [C] // NUSOD. 2022 International Conference on Numerical Simulation of Optoelectronic Devices. Turin: IEEE, 2022: 25-26.

[17] SHARMA V, SINGH S. 340 Gb/s All-Optical NOT, XOR, XNOR, $A \oplus B$, AB , OR and NOR Logic Gates Based on Cross-Gain Modulation in Semiconductor Optical Amplifiers [J]. Optik, 2021, 247: 168016.

- [18] DONG Jianji, ZHANG Xinliang, HUANG Dexiu. A Proposal for Two-Input Arbitrary Boolean Logic Gates Using Single Semiconductor Optical Amplifier by Picosecond Pulse Injection [J]. Optics Express, 2009, 17(10):7725-7730.
- [19] YANG Jingsheng, HAN Liuyan, ZHANG Hanyi, et al. Function-Lock Strategy in OR/NOR Optical Logic Gates based on Cross-Polarization Modulation Effect in Semiconductor Optical Amplifier [J]. Chinese Optics Letters, 2007, 5(10):566-568.
- [20] DENG Ning, CHAN Kit, CHAN Chunkit, et al. An All-Optical XOR Logic Gate for High-Speed RZ-DPSK Signals by FWM in Semiconductor Optical Amplifier [J]. IEEE Journal of Selected Topics in Quantum Electronics, 2006, 12(4):702-707.
- [21] TEJPAL R, KUMAR A, KUMAR R. Optimization of All-Optical Logic Gates Based on HNLF-MZI and Phase Shifter Configuration [C]// CSDP. International Conference on Signal & Data Processing. Singapore: Springer, 2022:81-93.
- [22] LOVKESH, SHARMA V, SINGH S. The Design of a Reconfigurable All-Optical Logic Device based on Cross-Phase Modulation in a Highly Nonlinear Fiber [J]. Journal of Computational Electronics, 2021, 20(1):397-408.
- [23] LIU Tonghui, JIA Dongfang, YANG Tianxin, et al. Experimental Validation of FWM Efficiency Improvement in Multichannel Fiber Laser by Using DF-HNLF [J]. Optics Communications, 2016, 367:12-16.
- [24] LAI D, KWOK CH, WONG K. All-Optical Picoseconds Logic Gates based on a Fiber Optical Parametric Amplifier [J]. Optics Express, 2008, 16(22):18362-18370.
- [25] YU Changyuan, CHRISTEN L, LUO Ting, et al. All-Optical XOR Gate Using Polarization Rotation in Single Highly Nonlinear Fiber [J]. IEEE Photonics Technology Letters, 2005, 17(6):1232-1234.
- [26] WU Wei, SUN Qibing, WANG Leiran, et al. Influence of Two-photon Absorption and Free-Carrier Effects on All-Optical Logic Gates in Silicon Waveguides [J]. Applied Physics Express, 2019, 12(4):042005.
- [27] GAO Shiming, WANG Xiaoyan, XIE Yanqiao, et al. Reconfigurable Dual-Channel All-Optical Logic Gate in a Silicon Waveguide Using Polarization Encoding [J]. Optics Letters, 2015, 40(7):1448-1451.
- [28] KHORASANINEJAD Mohammadreza, SAINI S. All-optical Logic Gates Using Nonlinear Effects in Silicon-on-insulator Waveguides [J]. Applied Optics, 2009, 48(25):F31-F36.
- [29] PASSARO V, DE PASSARO F. All-optical AND Gate Based on Raman Effect in Silicon-on-insulator Waveguide [J]. Optical and Quantum Electronics, 2006, 38:877-888.
- [30] WANG Jian, SUN Junqiang, ZHANG Xinliang, et al. Ultrafast All-Optical Three-Input Boolean XOR Operation for Differential Phase-Shift Keying Signals Using Periodically Poled Lithium Niobate [J]. Optics Letters, 2008, 33(13):1419-1421.
- [31] WANG Jian, SUN Junqiang, SUN Qizhen, et al. Experimental Observation of All-Optical Non-Return-to-Zero-to-Return-to-Zero Format Conversion

- Based on Cascaded Second-order Nonlinearity Assisted by Active Mode-locking [J]. Optics Letters, 2007, 32 (16) : 2462-2464.
- [32] WANG Jian, SUN Junqiang, SUN qizhen. Single-PPLN-based Simultaneous Half-adder, Half-subtractor, and OR Logic Gate: Proposal and Simulation [J]. Optics Express, 2007, 15 (4) : 1690-1699.
- [33] JINDAL P, ABOU HOURAN Mohamad, GOYAL D, et al. A Review of Different Techniques used to Design Photonic Crystal-Based Logic Gates [J]. Optik, 2023, 280 : 170794.
- [34] DATTA T, SEN M. Raman Mediated Ultrafast All-Optical NOR Gate [J]. Applied Optics, 2020, 59 (21) : 6352-6359.
- [35] HUSKO C, VO T, CORCORAN B, et al. Ultracompact All-optical XOR Logic Gate in a Slow-Light Silicon Photonic Crystal Waveguide [J]. Optics Express, 2011, 19 (21) : 20681-20690.
- [36] NATELSON D. Towards the Ultimate Transistor [J]. Physics World, 2009, 22 (6) : 27.
- [37] HOU Zhibo, XIANG Guoyong, DONG Daoyi, et al. Realization of Mutually Unbiased Bases for a Qubit with Only One Wave Plate: Theory and Experiment [J]. Optics Express, 2015, 23 (8) : 10018-10031.
- [38] KNILL E, LAFLAMME R, MILBURN G. A Scheme for Efficient Quantum Computation With Linear Optics [J]. Nature, 2001, 409 (6816) : 46-52.
- [39] FIURÁŠEK J. Linear-Optics Quantum Toffoli and Fredkin Gates [J]. Physical Review A—Atomic, Molecular, and Optical Physics, 2006, 73 (6) : 062313.
- [40] DIVINCENZO D. The Physical Implementation of Quantum Computation [J]. Fortschritte der Physik; Progress of Physics, 2000, 48 (9-11) : 771-783.
- [41] TANG Hao, SHANG Xiaowen, SHI Ziyu, et al. Simulating Photosynthetic Energy Transport on a Photonic Network [J]. NPJ Quantum Information, 2024, 10 (1) : 29.
- [42] HEBB D. The Organization of Behavior: A Neuropsychological Theory [M]. New York: Psychology Press, 2005.
- [43] 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.
- [44] BASSI F, BOTTI L, VERZEROLI L, et al. Parallelisation to Several Tens-of-thousands of Cores [J]. TILDA: Towards Industrial LES/DNS in Aeronautics: Paving the Way for Future Accurate CFD-Results of the H2020 Research Project TILDA, Funded by the European Union, 2015-2018, 2021 : 259-319.
- [45] LI Kebin. Energy-Efficient and High-Performance Processing of Large-scale Parallel Applications in Data Centers [M]. New York: Springer, 2015.
- [46] AATHILAKSHMI S, VAHEEN D, VIJAY H, et al. Analysis of RC Delay Using Multi Level Interconnects [C] // ICSSS. 2023 9th International Conference on Smart Structures and Systems. CHENNAI: IEEE, 2023 : 1-6.
- [47] DONG Po, QIAN Wei, LIAO Shirong, et al. Low loss Shallow-Ridge Silicon Waveguides [J].

- Optics Express, 2010, 18(14):14474-14479.
- [48] WANG Zhehui, XU Jiang, YANG Peng, et al. Alleviate Chip I/O Pin Constraints for Multicore Processors Through Optical Interconnects [C]// ASP-DAC. The 20th Asia and South Pacific Design Automation Conference. Chiba: IEEE, 2015:791-796.
- [49] LOPACINSKI L, EISSA M, GUTIERREZ J, et al. 5G and Beyond: Multi Baseband PSSS Architecture for 100 Gbps Wireless Communication [C]// MTTW. 2020 IEEE Microwave Theory and Techniques in Wireless Communications. Riga: IEEE, 2020:102-107.
- [50] WADE M, ANDERSON E, ARDALAN S, et al. An Error-Free 1 Tbps WDM Optical I/O Chiplet and Multi-Wavelength Multi-Port Laser [C]// OFC. Optical Fiber Communication Conference. San Francisco: IEEE, 2021:1-3.
- [51] IGARASHI K, TSURITANI T, MORITA I, et al. Ultra-Long-Haul High-Capacity Super-Nyquist-WDM Transmission Experiment Using Multi-core Fibers [J]. Journal of Lightwave Technology, 2015, 33(5):1027-1036.
- [52] NORBERG G, DEJANOVIC S, HESSELBOM H. Very High Density Interconnect Elastomer Chip Sockets [J]. IEEE Transactions on Advanced Packaging, 2006, 29(2):202-210.
- [53] HO Yingchieh, SU Chauchin. A 0.1-0.3 V 40-123 fJ/bit/ch On-chip Data Link With ISI-Suppressed Bootstrapped Repeaters [J]. IEEE Journal of Solid-state Circuits, 2012, 47(5):1242-1251.
- [54] SRIWAS S, CHANDRA D, KUMAR M. Long Distance Transmission in Optical-IDMA System [J]. SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology, 2023, 15(01):83-85.
- [55] KIKITSU A, KUROSAKI Y, SHIROTORI S, et al. Metallic/magnetic Multilayer for Wide-band Direct-on-chip EMI Shielding [C]// EMC SI. 2022 IEEE International Symposium on Electromagnetic Compatibility & Signal/Power Integrity. Spokane: IEEE, 2022:643-644.
- [56] ZHAO Miao, WEN Jing, HU Qiao, et al. A 3D Nanoscale Optical Disk Memory With Petabit Capacity [J]. Nature, 2024, 626(8000):772-778.
- [57] LAMON Simone, ZHANG Qiming, YU Haoyi, et al. Neuromorphic Optical Data Storage Enabled by Nanophotonics: A Perspective [J]. ACS Photonics, 2024, 11(3):874-891.
- [58] MCCULLOCH W, PITTS W. A Logical Calculus of the Ideas Immanent in Nervous Activity [J]. The Bulletin of Mathematical Biophysics, 1943, 5:115-133.
- [59] ROSENBLATT F. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain [J]. Psychological Review, 1958, 65(6):386.
- [60] RUMELHART D, HINTON G, WILLIAMS R. Learning Representations by Back-Propagating Errors [J]. Nature, 1986, 323(6088):533-536.
- [61] COVER T. Geometrical and Statistical Properties of Systems of Linear Inequalities With Applications in Pattern Recognition [J]. IEEE Transactions on Electronic Computers, 1965, (3):326-334.
- [62] ELMAN J. Finding Structure in Time [J].

- Cognitive Science,1990,14(2):179-211.
- [63] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-Based Learning Applied to Document Recognition[J]. Proceedings of the IEEE,1998, 86(11):2278-2324.
- [64] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is All You Need[C]// LUXBURG U V, GUYON I, BENGIO S, et al. NIPS' 17: Proceedings of the 31st International Conference on Neural Information Processing Systems. New York:Curran Associates Inc.,2017: 6000-6010.
- [65] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale[EB/OL]. [2021-06-03]. <https://arxiv.org/abs/2010.11929>.
- [66] HE Kaiming, CHEN Xinlei, XIE Saining, et al. Masked Autoencoders are Scalable Vision Learners[C]// CVPR. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 16000-16009.
- [67] DEVLIN J, CHANG Mingwei, LEE K, et al. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding [EB/OL]. (2024-05-24). <https://arxiv.org/abs/1810.04805>.
- [68] MAASS W. Networks of Spiking Neurons: The Third Generation of Neural Network Models[J]. Neural Networks,1997,10(9):1659-1671.
- [69] SHEN Yiwei, LI Ruiqian, LIU Guanting, et al. Deep Photonic Reservoir Computing Recurrent Network[J]. Optica,2023,10(12):1745-1751.
- [70] LIU Keqin, ZHANG Teng, DANG Bingjie, et al. An Optoelectronic Synapse Based on α -In₂Se₃ With Controllable Temporal Dynamics for Multimode and Multiscale Reservoir Computing [J]. Nature Electronics,2022,5(11):761-773.
- [71] FELDMANN J, YOUNGBLOOD N, KARPOV M, et al. Parallel Convolutional Processing Using an Integrated Photonic Tensor Core [J]. Nature, 2021,589(7840):52-58.
- [72] FELDMANN J, YOUNGBLOOD N, WRIGHT C, et al. All-optical Spiking Neurosynaptic Networks With Self-learning Capabilities [J]. Nature, 2019,569(7755):208-214.
- [73] LIN Xing, RIVENSON Y, YARDIMCI N, et al. All-Optical Machine Learning Using Diffractive Deep Neural Networks [J]. Science, 2018, 361(6406):1004-1008.
- [74] SHEN Yichen, HARRIS N, SKIRLO S, et al. Deep Learning with Coherent Nanophotonic Circuits[J]. Nature Photonics, 2017, 11(7): 441-446.
- [75] CHENG Zengguang, RÍOS C, PERNICE W, et al. On-Chip Photonic Synapse [J]. Science Advances,2017,3(9):e1700160.
- [76] BOGAERTS W, PÉREZ D, CAPMANY J, et al. Programmable Photonic Circuits [J]. Nature, 2020,586(7828):207-216.
- [77] HARRIS N, CAROLAN J, BUNANDAR D, et al. Linear Programmable Nanophotonic Processors [J]. Optica,2018,5(12):1623-1631.
- [78] WEI Maoliang, XU Kai, TANG Bo, et al. “Zero Change” Platform for Monolithic Back-End-of-Line Integration of Phase Change Materials in Silicon Photonics [EB/OL]. (2023-08-30). <https://arxiv.org/abs/2308.15723>.
- [79] XU Peipeng, ZHENG Jiajiu, DOYLEND J, et al.

- Non-Volatile Integrated-silicon-photonic Switches Using Phase-change Materials [C]// ACPC. Asia Communications and Photonics Conference. Chengdu: Optica Publishing Group, 2019: S4H. 6.
- [80] TAIT A, DE LIMA T, NAHMIAS M, et al. Continuous Calibration of Microring Weights for Analog Optical Networks [J]. IEEE Photonics Technology Letters, 2016, 28(8): 887-890.
- [81] TAIT A, WU Allie X, DE LIMA T, et al. Microring Weight Banks [J]. IEEE Journal of Selected Topics in Quantum Electronics, 2016, 22(6): 312-325.
- [82] DEL' HAYE P, SCHLIESSER A, ARCIZET O, et al. Optical Frequency Comb Generation from a Monolithic Microresonator [J]. Nature, 2007, 450(7173): 1214-1217.
- [83] BAI Bowen, YANG Qipeng, SHU Haowen, et al. Microcomb-based Integrated Photonic Processing Unit [J]. Nature Communications, 2023, 14(1): 66.
- [84] SHU Haowen, CHANG Lin, TAO Yuansheng, et al. Microcomb-driven Silicon Photonic Systems [J]. Nature, 2022, 605(7910): 457-463.
- [85] APPELTANT L, SORIANO M, VAN DER SANDE G, et al. Information Processing Using a Single Dynamical Node as Complex System [J]. Nature Communications, 2011, 2(1): 468.
- [86] NAKAJIMA M, INOUE K, TANAKA K, et al. Physical Deep Learning with Biologically Inspired Training Method: Gradient-free Approach for Physical Hardware [J]. Nature Communications, 2022, 13(1): 7847.
- [87] FREIBERGER M, SACKESYN S, MA Chonghuai, et al. Improving Time Series Recognition and Prediction With Networks and Ensembles of Passive Photonic Reservoirs [J]. IEEE Journal of Selected Topics in Quantum Electronics, 2019, 26(1): 1-11.
- [88] WU Lin, ZHANG Ziyang. Only-Train-Electrical-to-Optical-Conversion (OTEOC): Simple Diffractive Neural Networks With Optical Readout [J]. Optics Express, 2022, 30(15): 28024-28037.
- [89] LOU Minhan, LI Yingjie, YU Cunxi, et al. Effects of Interlayer Reflection and Interpixel Interaction in Diffractive Optical Neural Networks [J]. Optics Letters, 2023, 48(2): 219-222.
- [90] IZHIKEVICH E. Dynamical Systems in Neuroscience [M]. Massachusetts: MIT Press, 2007.
- [91] KIM Chanju, YE Chaochao, ZHENG Yi, et al. Design and Fabrication of AlGaAs-on-Insulator Microring Resonators for Nonlinear Photonics [J]. IEEE Journal of Selected Topics in Quantum Electronics, 2022, 29(1: Nonlinear Integrated Photonics): 1-14.
- [92] VIEU C, CARCENAC F, PEPIN A, et al. Electron Beam Lithography: Resolution Limits and Applications [J]. Applied Surface Science, 2000, 164(1-4): 111-117.
- [93] YE Chaochao. Aluminum Gallium Arsenide-on-Insulator Integrated Nonlinear Photonics [D]. Copenhagen: Technical University of Denmark, 2024.
- [94] RAMBU A, APETREI A, TASCU S. Role of the High Vacuum in the Precise Control of Index Contrasts and Index Profiles of LiNbO3 Waveguides Fabricated by High Vacuum Proton

- Exchange[J]. Optics & Laser Technology, 2019, 118:109-114.
- [95] ZHU Di, SHAO Linbo, YU Mengjie, et al. Integrated Photonics on Thin-Film Lithium Niobate[J]. Advances in Optics and Photonics, 2021, 13(2):242-352.
- [96] LU H, SADANI B, COURJAL N, et al. Enhanced Electro-Optical Lithium Niobate Photonic Crystal Wire Waveguide on a Smart-Cut Thin Film[J]. Optics Express, 2012, 20(3):2974-2981.
- [97] DE MICHELI M, BOTINEAU J, NEVEU S, et al. Extension of Second-harmonic Phase-matching Range in Lithium Niobate Guides [J]. Optics Letters, 1983, 8(2):116-118.
- [98] RAO A, ABDELSALAM K, SJAARDEMA T, et al. Actively-Monitored Periodic-poling in Thin-Film Lithium Niobate Photonic Waveguides With Ultrahigh Nonlinear Conversion Efficiency of 4600% $\text{W}^{-1}\text{cm}^{-2}$ [J]. Optics Express, 2019, 27(18):25920-25930.
- [99] KOJIMA A, TESHIMA K, SHIRAI Y, et al. Organometal Halide Perovskites as Visible-light Sensitizers for Photovoltaic Cells[J]. Journal of the American Chemical Society, 2009, 131(17):6050-6051.
- [100] WALTERS G, SUTHERLAND B, HOOGLAND S, et al. Two-photon Absorption in Organometallic Bromide Perovskites [J]. ACS Nano, 2015, 9(9):9340-9346.
- [101] LI Ruxue, YU Jiahao, WANG Sheng, et al. Surface Modification of All-inorganic Halide Perovskite Nanorods by a Microscale Hydrophobic Zeolite for Stable and Sensitive Laser Humidity Sensing[J]. Nanoscale, 2020, 12(25):13360-13367.
- [102] SHANG Xinxin, GUO Linguang, ZHANG Huanian, et al. Titanium Disulfide Based Saturable Absorber for Generating Passively Mode-locked and Q-switched Ultra-fast Fiber Lasers [J]. Nanomaterials, 2020, 10(10):1922.
- [103] WU Jiayang, YANG Yunyi, QU Yang, et al. 2D Layered Graphene Oxide Films Integrated with Micro-ring Resonators for Enhanced Nonlinear Optics[J]. Small, 2020, 16(16):1906563.
- [104] MOUCHEL P, KEMEL M, SEMAAN G, et al. Limitations of Graphene Nanocoated Optical Tapers for High-power Nonlinear Applications [J]. Optical Materials: X, 2019, 1:100003.
- [105] NOBILE N, ERICKSON J, RÍOS C, et al. Time-Resolved Temperature Mapping Leveraging the Strong Thermo-Optic Effect in Phase-Change Materials[J]. ACS Photonics, 2023, 10(10):3576-3585.
- [106] PITCHAPPA P, KUMAR A, PRAKASH S, et al. Volatile Ultrafast Switching at Multilevel Nonvolatile States of Phase Change Material for Active Flexible Terahertz Metadevices [J]. Advanced Functional Materials, 2021, 31(17):2100200.
- [107] PELGRIN V, YOON H, CASSAN E, et al. Hybrid Integration of 2D Materials for On-chip Nonlinear Photonics [J]. Light: Advanced Manufacturing, 2023, 4(3):311-333.
- [108] CHANG Lin, PFEIFFER M, VOLET N, et al. Heterogeneous Integration of Lithium Niobate and Silicon Nitride Waveguides for Wafer-scale Photonic Integrated Circuits on Silicon [J].

Optics Letters,2017,42(4):803-806.

[109] XU Zhihao, ZHOU Tiankuang, MA Muzhou, et al. Large-scale Photonic Chiplet Taichi Empowers 160-TOPS/W Artificial General Intelligence [J]. Science, 2024, 384 (6692) : 202-209.

[110] XUE Zhiwei, ZHOU Tiankuang, XU Zhihao, et al. Fully Forward Mode Training for Optical Neural Networks [J]. Nature, 2024, 632 (8024) :280-286.

[111] ZHAO Yun, CHEN Hang, LIN Min, et al. Optical Neural Ordinary Differential Equations [J]. Optics Letters,2023,48(3):628-631.

[112] SUN Ke, TAN Dezhi, FANG Xinyuan, et al. Three-Dimensional Direct Lithography of Stable Perovskite Nanocrystals in Glass [J]. Science, 2022,375(6578):307-310.

[113] XU Xiaoyi, WANG Tianxin, CHEN Pengcheng, et al. Femtosecond Laser Writing of Lithium Niobate Ferroelectric Nanodomains [J]. Nature, 2022,609(7927):496-501.

作者贡献说明

谢 斌:搜集、整理资料,撰写稿件;
谢 红:设计文章框架,梳理资料,修改稿件。