中国激光

# 光电智能计算研究进展与挑战

成骏伟[1]，江雪怡[1]，周海龙[1]，董建绩[1,2*]

[1]华中科技大学武汉光电国家研究中心，湖北 武汉 430074；

[2]湖北光谷实验室，湖北 武汉 430074

**摘要** 随着人工智能技术的高速发展，全球的计算量急剧增长，需要以快速、高效的方式处理海量数据，这对计算硬件的算力和能效提出了较高的要求。受限于电子器件的固有极限和冯·诺依曼架构，传统的电子计算在速度和能效方面遇到了难以突破的瓶颈。光电智能计算充分融合光学的多维复用、大带宽、低能耗等优势和电学的细粒度灵活控制特性，具有光算电控和软硬协同的特点，是一种更实用、更有竞争力的人工智能计算加速方案。回顾了光电智能计算的研究进展，探讨了目前用于光学信号处理和光学神经网络的主流计算架构在线训练算法以及算力、能效提升方面的挑战，并进行了展望。

**关键词** 光计算；光电智能计算；人工智能；计算加速；光学信号处理；光学神经网络

**中图分类号** TN45 **文献标志码** A **DOI**：10.3788/CJL202249.1219001

## 1 引 言

人工智能是目前最活跃的研究领域之一，其中以人工神经网络为代表的人工智能模型是一种模拟大脑神经突触网络的计算模型，已经被广泛用于计算机视觉、语音识别、自动驾驶等领域[1-4]。在近十年里，人工智能技术经历了爆炸式的增长，全球的计算量急剧增长，迫切需要以快速、高效的方式处理海量数据。然而，基于冯·诺依曼架构的电子计算机的局限性日渐凸显，由于处理器和存储器是物理分离的，数据需要在处理器和存储器之间来回传输，处理器的运算速度和内存访问速度的不匹配导致的带宽瓶颈和功耗瓶颈严重限制了冯·诺依曼架构的计算能效，即“内存墙”问题。此外，随着摩尔定律的放缓甚至失效，集成电路的晶体管数量正在逼近物理极限，电子微处理器的时钟速率被限制在几个GHz[5-7]，这已经不能满足超高速、低延迟的海量数据处理的需求，学术界和工业界都在致力于开发新的硬件架构和软件算法，力求在提升算力的同时，降低运算的能耗，以满足人工神经网络和深度学习的应用要求。

目前，集成电路芯片在人工智能计算加速领域占据主要地位，例如中央处理器（CPU）、图形处理器（GPU）、专用集成电路（ASIC）和现场可编程门阵列（FPGA）等是目前主流的神经网络训练与推理的计算硬件。在神经形态电子学的概念[8]被提出之后，国内外已成功研制了多款神经形态芯片[9-10]。与此同时，全球科技巨头企业也迅速布局，IBM TrueNorth[11]、Google TPU[12]、Intel Loihi[13]等人工智能计算加速芯片先后发布，在提高训练与推理任务的速度和能效方面取得了重要进展。然而，神经形态电子芯片的一个主要限制是互连密度，因此目前神经形态处理频率基本被限制在MHz量级[14]。受限于集成电路中的串扰、能耗、带宽和时延等，仅通过提高集成密度和工作频率来进一步提高微电子芯片的信息处理能力已经难以为继。此外，电互连中的带宽和互连密度难以满足高性能计算的需求[15]，迫切需要新的计算硬件来打破传统微电子计算框架的固有限制。

光电智能计算用光子代替电子进行计算，因此可以通过克服电子固有的局限性而显著提高计算速度和能效。光子器件在进行模拟运算时，相比电子

**收稿日期**：2022-02-14；**修回日期**：2022-04-07；**录用日期**：2022-04-24

**基金项目**：国家自然科学基金（62075075，U21A20511）、湖北光谷实验室创新科研项目（OVL2021BG001）

**通信作者**：*jjdong@hust.edu.cn

器件有以下几个方面的优势：1)光具有波长、波导模式、相位、振幅和偏振等多个物理维度资源，具有天然的并行计算能力，可以成倍地提高算力；2)电信号在金属导线中传输必然伴随热量的产生，而光在计算时没有欧姆加热，因此其能耗远小于电子计算；3)与电子器件相比，光子器件有更大的带宽，因此在宽带模拟信号的处理上远胜于电子器件；4)电子器件具有电阻电容(RC)效应，因此会在计算中产生延时，而光学器件则基本没有延迟，具有更快的响应时间；5)光子的物理本质是玻色子，这意味着它们可以在没有相互干扰的情况下传输，而电子的本质是费米子，较容易被扰动。这些优异特性使得光电智能计算特别适用于实现包含大量神经元和突触的大规模神经网络，基于光电智能计算的光学神经网络技术应运而生[16]。目前，由常规光子器件构建的神经形态计算硬件，如马赫-曾德尔干涉仪(MZI)网络和微环谐振器(MRR)阵列，每次乘加(MAC)的能耗大概在飞焦水平，这项指标比最先进的互补金属氧化物半导体(CMOS)计算硬件小两个数量级，说明光学神经网络在实现超高速计算的同时，在计算能效上也远优于电学神经网络，在高并发、高吞吐量、计算密集型的超算平台和数据中心等应用场景中天然地具有显著优势。光互连已经广泛用于数据中心，且大规模互连的时间和能源消耗显著降低[17]，光电智能计算和光互连技术的结合将为面向超算平台和数据中心需求的专用计算硬件提供创新机遇。此外，深度学习技术也被广泛应用于计算光学成像中，可获得场景目标的高质量图像与高维度物理信息[18-19]。

算力、能效、算法是人工智能计算至关重要的三个方面，硬件友好的算法和计算硬件的算力、能效成为了人工智能发展的关键[20-21]。光电智能计算已经经过近十年的飞速发展，但是关于光电智能计算领域中的算力、能效和算法的系统梳理还比较欠缺。本文首先回顾了光电智能计算领域的发展历史和研究进展，然后对算力、能效和算法进行了系统梳理，并针对在线训练算法和计算性能指标提升等现阶段的重要挑战进行了探讨，最后对光电智能计算的发展趋势进行了展望。

## 2 光电智能计算的研究进展

近年来，光电智能计算领域的研究得到了蓬勃发展，矩阵矢量乘法(MVM)作为最基本的矩阵计算之一，在现代信号处理和人工智能算法中占用了大部分计算开销。目前主流的光学MVM有三种类型，如图1所示，分别是基于平面光转换(PLC)的矩阵计算、基于MZI的矩阵计算和基于波分复用(WDM)的矩阵计算。这三类光学MVM的基本工作原理可以参见文献[22]，该文献为光学矩阵计算的原理提供了系统全面的论述，文献[23]则详细介绍了光学MVM运算在光学信号处理和人工智能领域中的发展动态。

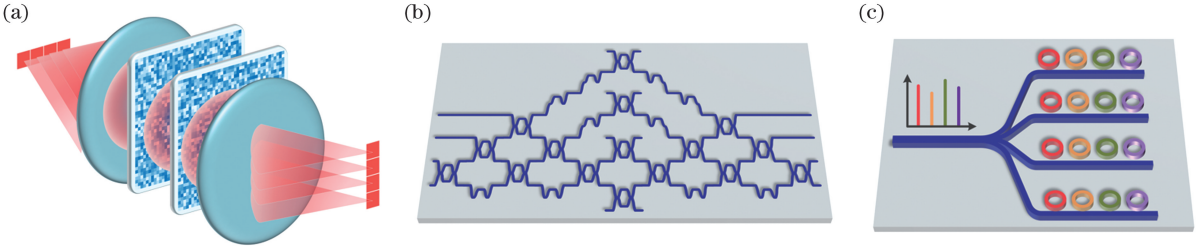


图1 光学MVM的类型。(a)基于PLC的光学MVM；(b)基于MZI的光学MVM；(c)基于WDM的光学MVM

Fig. 1 Categories of photonic MVM. (a) PLC-based photonic MVM; (b) MZI-based photonic MVM; (c) WDM-based photonic MVM

基于PLC的光学MVM是通过光在自由空间的衍射实现的。起初，只有一些固定的矩阵计算可以通过光学实现，如傅里叶变换[24]。最早的可编程光学MVM是用基于单平面光转换(SPLC)[25]的空间光学元件演示的，随后提出的多平面光转换(MPLC)架构[26]使得输入、输出向量分布在整个二维平面上，极大地拓展了运算规模，MPLC架构因此被广泛应用于全光机器学习[27-29]、光子伊辛机[30-31]和神经形态光电计算[32]等领域。但是MPLC架构仍受限于数据刷新速率和空间带宽积，而且空间光学元件占地面积较大。2017年，Tang等[33]提出了一种基于多模干扰耦合器的新型集成可重构单元光模式转换器，其随后被应用于全光片上多输入多输出(MIMO)模式解复用[34]，是MPLC架构的集成化实现。2020年，Saygin等[35]提出了一种更通用的集成MPLC架构并对其鲁棒性进行了深入分析，并在硅光芯片上进行了实验验证[36]，这在一定程度上弥补了MPLC架构难以集成化的

缺陷。

随着集成光电子学的发展，以MZI网络为代表的片上光电智能计算架构被提出，其通过MZI网络进行模拟相干运算，具有可扩展性强、能效高和可编程性强等优点。1994年，Reck等[37]提出了一种三角分解方法，可以将任意矩阵分解为一系列的二维变换矩阵，证明了由分束器和相移器构建的三角形网络能够实现任意有限维度的酉矩阵，这为基于MZI网络的光电智能计算奠定了基础。2016年，Clements等[38]提出了矩形分解方法，只需要三角分解方法的一半的光学深度，而且光学损耗也显著降低。此外，Shokraneh等[39]还提出了一种菱形排布的MZI网络，与三角形网络相比，菱形网络对相位误差和插入损耗有更高的鲁棒性。2017年，Shen等[40]研制了基于三角形级联MZI网络的光电计算芯片，通过结合光域的线性运算和电域的非线性激活函数，构建了全连接神经网络，成功实现了元音识别。2021年，Xu等[41]提出了一种硅基光学相干点积芯片，并成功地演示了基于AUTOMAP神经网络模型的图像重建，重建的图像质量与32 bit数字计算机相当。除神经网络外，MZI网络还可以实现偏振处理、网页排名、MIMO解扰和光子伊辛机等应用。本研究团队一直致力于片上光电智能计算架构和算法的研究，聚焦光学信号处理和人工智能计算领域的关键问题，取得了多功能片上偏振处理器[42]、Google PageRank光子计算加速器[43]以及自配置、完全可重构的硅光信号处理器[44]等重要进展，拓展了MZI网络的应用场景。2020年，Roques-Carmes等[45]提出了光子递归伊辛采样器(PRIS)，这是一种为并行架构定制的启发式方法，可从任意伊辛问题的分布中进行快速有效的采样，并实验验证了PRIS[46]。

此外，由于WDM技术的日益成熟，光的波长资源得到了有效的利用，通过将单波长独立调制与多波长并行处理相结合，可以构建高并行性的计算硬件。基于WDM的光学MVM是一种非相干矩阵计算方法。2011年，Xu等[47]首次提出用于矩阵运算的硅基MRR阵列概念，Yang等[48]于次年实验演示了使用硅基MRR阵列进行矩阵乘法的实验结果。普林斯顿大学的Prucnal研究团队首次提出基于MRR阵列的Broadcast and Weight架构[49]，并在光学神经网络领域开展了一系列相关的工作[50-55]。本研究团队巧妙地结合了MZI和MRR结构的特点，提出了一种MZI辅助的MRR(MZI-MRR)结构[56]，并基于MZI-MRR阵列实现了可编程脉冲处理器[57]和用于图像处理的光学卷积加速器[58]。2019年，Feldmann等[59]通过引入相变材料(PCM)，提出了一种基于WDM的交叉开关阵列的光学脉冲神经网络，并实现了字母识别的应用。2020年，Miscuglio等[60]通过引入张量计算的概念，提出了一种基于MRR阵列的4 bit光子张量核，基于并行数据处理，MAC运算的速度加快了数倍，同时系统的能耗和延迟得到显著降低。此外，目前的孤子微腔光频梳可以产生数十条频率间隔恒定的梳状频率线，作为WDM架构的多波长光源，可以充分发挥光在波长维度的潜力。2021年，Feldmann等[61]提出了一种基于孤子微腔光频梳和PCM阵列的集成光子张量核，每秒可运行的MAC数可达$2\times10^{12}$。同年，Xu等[62]提出的通用光学矢量卷积加速器使用了一个能够产生90个梳齿的孤子微腔光频梳，凭借超高的计算并行性，实现了高达$11\times10^{12}$ operation/s的算力。

光子智能计算领域方兴未艾，国内外的研究机构经过系统深入的研究，在光学神经网络和光学信号处理等方面均取得了丰硕的成果。图2总结了目前已经实现的全连接神经网络[40]、卷积神经网络[62-65]、脉冲神经网络[49, 59]、储备池计算[66-68]、非线性激活函数[69-71]等光学神经网络计算以及模式解扰[72]、模式分析[73]、光学加密和感知[74-75]、偏振分析[42]、频谱整形[76]等智能光学信号处理应用。

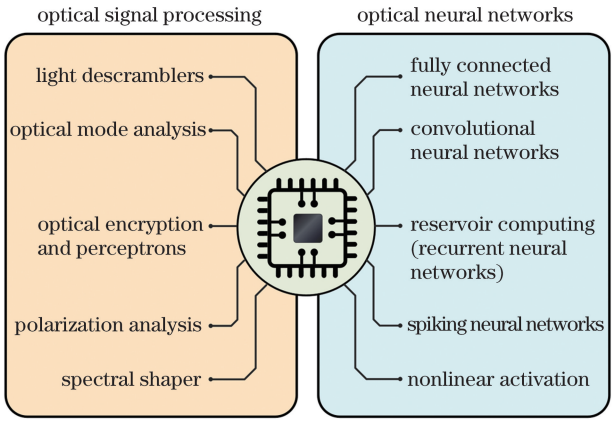


图2 光电智能计算的应用总结

Fig. 2 Summary of applications in optoelectronic intelligent computing

## 3 光电智能计算面临的挑战

光电智能计算充分融合光学的多维复用、大带宽、低能耗等优势和电学的细粒度灵活控制特性，具

有“光算电控”和“软硬协同”的特点，通过充分发挥光电各自的优势，可以显著提高人工智能计算的速度和能效。然而，目前光电智能计算的在线训练算法以及算力、能效的提升仍然存在诸多重要挑战。这些挑战制约着光电智能计算的发展和产业化，但同时也带来了创新和突破的机遇。

### 3.1 在线训练算法

人工智能算法可分为训练和推理两部分，但由于光子难以被存储，且无法直接获取光子的状态，因此目前多数光电智能计算芯片都采用“电域训练，光域推理”的方法实现人工智能算法，即在电子计算机上对神经网络的仿真模型进行训练，再将训练出的模型参数加载到光子芯片上进行推理。然而，这种离线训练明显存在数值仿真模型与实际物理模型之间的差异，更重要的是，如果光子计算芯片神经网络的实现始终依赖于电域的训练，那么光子芯片所具备的低延迟、高能效等优势则无法充分发挥。因此，开发硬件友好的光电智能计算在线训练算法具有重要意义。下面分别介绍面向光学神经网络芯片和衍射光学神经网络的光学方法实现的反向传播（BP）算法，以及面向片上智能光信号处理和光电智能计算的随机梯度下降（SGD）算法。

在神经网络中，学习的本质是神经网络参数的调优，学习过程由信号的正向传播与误差的反向传播两个过程组成。神经网络中的权重参数一般采用随机梯度下降法进行训练，计算每个参数梯度的最常用方法是基于链式法则的反向传播，权重参数通过BP算法进行迭代更新。2018年，Hughes等[77]基于伴随变量法实现了BP算法的光子模拟，从理论上证明了BP算法可以在基于光学干涉计算单元（OIUs）的光学神经网络芯片中实现，其算法流程如图3(a)所示。通过原始光场、伴随光场和干涉光场在传播中的光场分布，可以得到代价函数相对于相移器向收敛方向下降的梯度值，进而计算下一轮迭代中各相移器的相位配置，最终芯片逐步收敛至最优配置。这种梯度测量方法在无损系统中是精确的，并且Hughes等[77]通过数值仿真证明了这种梯度测量方法对具有模式相关损耗的系统也是有效的。此外，这种在线训练方法可以在恒定时间内计算任意数量可调参数下的代价函数梯度，为未来更大规模的片上混合光电网络的在线训练奠定了基础。

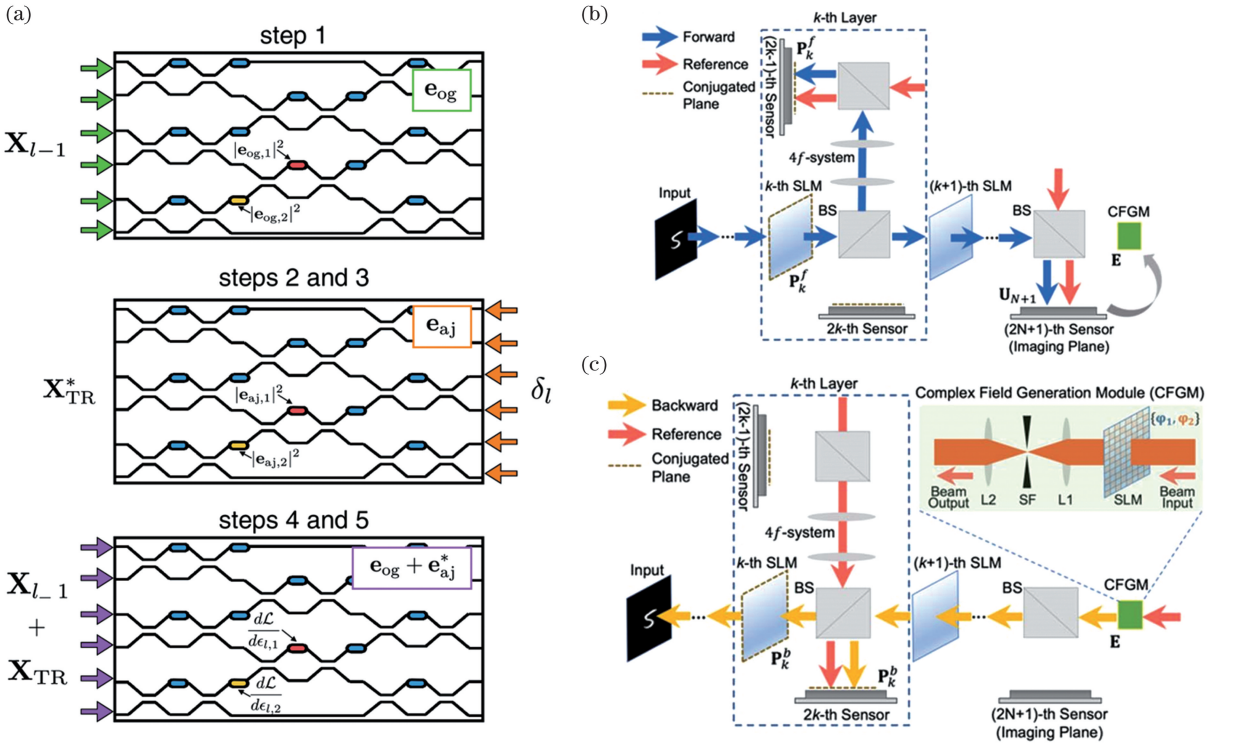


图3 基于BP算法在线训练光学神经网络。(a)集成光学神经网络芯片[77]；(b)衍射光学神经网络的前向传播[78]；(c)衍射光学神经网络的反向传播[78]

Fig. 3 *In situ* training of optical neural networks through BP algorithm. (a) Chip of integrated optical neural networks[77]; (b) forward propagation of diffractive optical neural networks[78]; (c) backward propagation of diffractive optical neural networks[78]

2020年，清华大学的研究团队提出了一种衍射光学神经网络的在线训练方法，将BP算法光学化，实现对线性和非线性衍射光学神经网络的在线训练，提高了核心计算模块的训练速度和能量效率[78]。衍射光学神经网络结构在物理上是由级联空间光调制器（SLMs）实现的，可以通过编程SLMs对网络的衍射调制系数进行配置以匹配特定的任务需求。衍射调制系数的每次迭代训练包括前向传播、误差计算、反向传播和梯度更新四个步骤，其前向传播和反向传播流程分别如图3（b）、（c）所示。通过测量前后传播光场，可以精确地计算出衍射层相对于权值的代价函数梯度。通过编程SLMs来更新衍射调制权值，以最小化预测和目标输出之间的误差，并以光速执行推理任务。

本研究团队主要聚焦于片上在线训练算法的研究，针对光子芯片的片上MZI网络架构的隐式计算特点，提出了一种面向光子芯片的SGD算法，其简要流程如图4（a）所示，其中 $C_{\mathrm{temp}}$ 是临时代价函数值，$C_i$ 是迭代次数为 $i$ 时的代价函数值，$C_{\mathrm{target}}$ 是目标代价函数值。片上训练的具体流程是，首先设置训练的目标，然后依次微调片上相移器的状态，并计算实时的代价函数（CF），在训练的迭代过程中，如果CF提升，则更新参数，否则就恢复原来的参数，直至达到训练目标或迭代次数达到预设值。在此基础上，逐步建立了“光算电控-软硬协同”的智能光电计算体系，取得了一系列研究进展，包括多功能片上偏振处理器[42]、Google PageRank光子计算加速器[43]以及自配置、完全可重构的硅光信号处理器[44]，分别如图4（b）～（d）所示。在SGD算法的实际应用中，根据不同的功能需求灵活设计代价函数，实时检测各端口的信号质量，在迭代过程中芯片配置逐步收敛至最优。图4（d）展示的是自配置、完全可重构的硅光信号处理器的代价函数和芯片功能在训练过程中的优化情况。通过使用由深度学习修正的SGD算法进行在线训练，该智能光子处理器可以在不知道芯片内部结构的情况下通过完全的自配置实现多通道光开关、MIMO解扰器和可调谐光滤波器三种典型光信号处理功能。

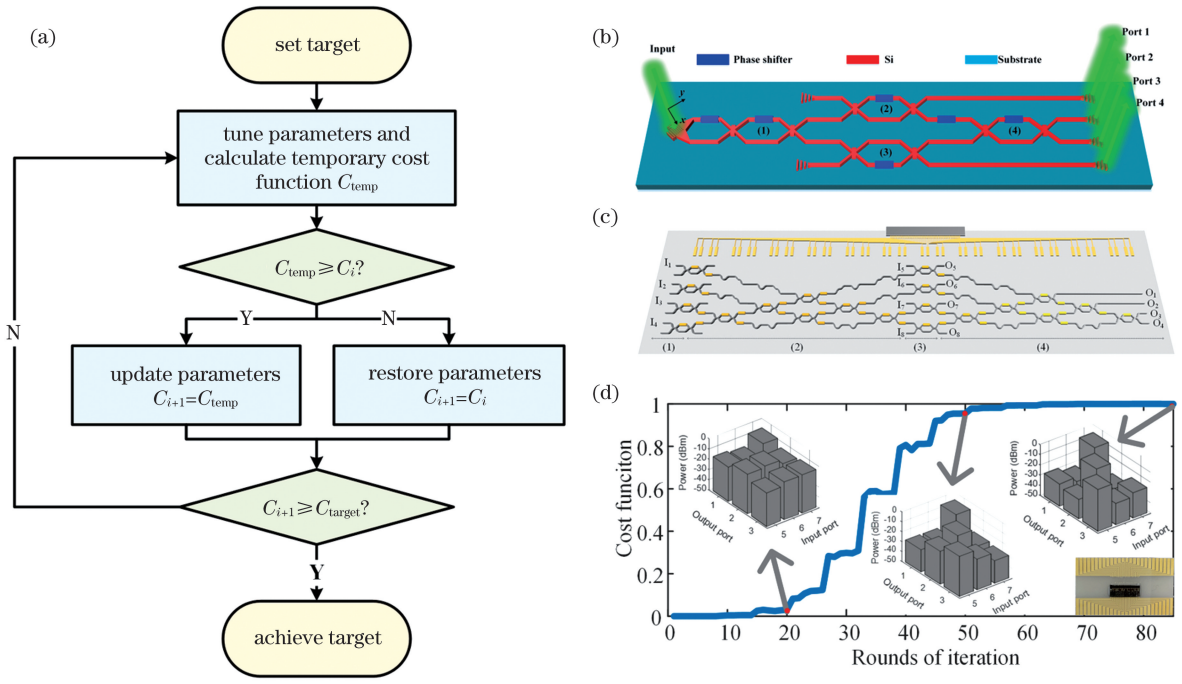


图4　通过SGD算法在线训练光电智能计算芯片。（a）SGD算法的简要流程；（b）多功能片上偏振处理器[42]；（c）用于Google PageRank算法的光子加速器[43]；（d）自配置、完全可重构的硅光信号处理器[44]

Fig. 4　Online training of optoelectronic intelligent computing chip through SGD algorithm. (a) Brief flow chart of SGD algorithm; (b) multifunctional on-chip polarization processor[42]; (c) photonic accelerator for Google PageRank algorithm[43]; (d) self-configuring and fully reconfigurable silicon photonic signal processor[44]

### 3.2　算力和能效的提升

光电智能计算肩负着突破传统微电子芯片算力和能效局限的重要任务，因此在关于光电智能计算的研究中，算力与能效成为了评价光电智能计算性能的重要指标。在讨论光电智能计算的算力和能效前，先考虑微电子芯片的算力与能效的计算方法，并

且将二者的影响因素进行比较。微电子芯片的算力公式为

$$s = c_o \times o_c \times c_s, \tag{1}$$

式中：$s$ 为芯片的算力；$c_o$ 为处理器核数；$o_c$ 为单个芯片每周期的浮点运算数；$c_s$ 为每秒周期数，与处理器的主频呈正相关。

与微电子芯片的处理器核数、每周期浮点运算数与处理器的主频分别对应的是光子智能计算芯片的并行数、单次运算规模以及波特率。考虑到这三者乘积的物理意义为光子芯片每秒可运行的 MAC 数，最终每秒钟的操作数需要乘以 2，即

$$p = 2 \times V \times M \times B, \tag{2}$$

式中：$p$ 为光子芯片每秒可运行的操作数；$V$ 为并行数；$M$ 为单次运算规模；$B$ 为波特率。能效的定义为算力与能耗的比值。与微电子芯片相比，光子芯片的能耗构成一般更为复杂，尤其是光电混合芯片，需要同时考虑光路和电路中的消耗以及光电转换过程中产生的损耗。

在光路中，为保证一定的信噪比，同时克服定向耦合器、交叉波导和非线性器件的能量损耗，光路中存在一个维持运算的最低光功率 $P_{\text{minlight}}$。$P_{\text{minlight}}$ 决定了一个全光的光子芯片的能耗下限。光电转换部分包括了激光器和探测器的损耗，其总效率记为 $\eta$。$P_{\text{minlight}}/\eta$ 即是光路的最低运行功率，驱动激光器所需的功率记为 $P_{\text{TX}}$。

消耗电能的器件可分为两部分，一部分直接作用于光路，包括调制器和光频梳，另一部分则不直接作用于光路，包括 DAC、ADC 和 TIA 等。这些器件的功率记为 $P_{\text{mod}}$、$P_{\text{comb}}$、$P_{\text{DAC}}$、$P_{\text{ADC}}$ 和 $P_{\text{TIA}}$，分别指调制器、光频梳、DAC、ADC 和 TIA 的功率。因此，光电智能计算芯片能耗的一个通用公式为

$$p_o = \max(P_{\text{minlight}}/\eta, P_{\text{TX}}) + P_{\text{mod}} + P_{\text{comb}} + P_{\text{DAC}} + P_{\text{ADC}} + P_{\text{TIA}}, \tag{3}$$

式中：$p_o$ 为光电智能计算芯片能耗。

目前的光子芯片矩阵规模较小，对于光电智能计算芯片，其在光路部分的功率可以忽略不计，只需考虑激光器运行产生的总功率，此时就可以得到

$$p_o = P_{\text{TX}} + P_{\text{mod}} + P_{\text{comb}} + P_{\text{DAC}} + P_{\text{ADC}} + P_{\text{TIA}}。 \tag{4}$$

表 1 展示了几种典型光电子芯片和微电子芯片的算力、能效对比，其中微电子芯片部分已换算成 8 bit 精度下的算力及能效。从表 1 可以看出，尽管已报道的光电子芯片的算力尚未超过微电子芯片，但算力最优的光电子芯片已经接近目前商用的微电子芯片。且表 1 中的光电子芯片算力数据均为单芯片算力，而微电子芯片由于封装工艺的成熟一般含有多个计算核心。在能效方面，光电子芯片每次 MAC 操作的能效普遍在 fJ 量级，而一般的微电子芯片每次 MAC 操作的能效则在 pJ 量级，差距在两个数量级以上。

表 1　微电子芯片与光电子芯片的算力和能效对比

Table 1　Comparison of computing capacity and energy efficiency of microelectronic chips and optoelectronic chips

| Technology | Computing capacity /($10^{12}$ operation/s) | Energy efficiency |
|---|---|---|
| Coherent MZI mesh[40] | 3.2 | 30 fJ per MAC |
| Time-wavelength interleaving photonic convolutional accelerator[62] | 11 | 0.39 fJ per MAC |
| Photonic WDM/PCM in-memory computing[61] | 4.302 | 17 fJ per MAC |
| Google TPU[79] | 23 | 0.43 pJ per MAC |
| NVIDIA Tesla T4[80] | 130 | 1.08 pJ per MAC |
| HUAWEI Ascend 310[81] | 16 | 1 pJ per MAC |
| HUAWEI Ascend 910[82] | 640 | 1.09 pJ per MAC |

图 5 总结了三种典型的光电智能计算架构。下面将梳理目前主流光电智能计算架构的算力和能效指标的评价方法，同时验证前文中提到的评价方法的普适性。

首先以三角形排布的 MZI 网络[40]为例，如图 5(a)所示，假设系统的数据检测频率为 100 GHz，系统包含 2 层，而每层包含 $N \times N$（$N=4$）权重矩阵。如果执行非线性激活的操作可忽略不计，则其算力为 $6.4\times10^{12}$ operation/s。根据文献[40]，假设可饱和吸收体的饱和功率为 $p'$，波导的横截面为 $A=0.2\ \mu\text{m}\times0.5\ \mu\text{m}$，假设光路中的传播损耗可以忽略不计，则光路中的最低能耗为 $p_o = N \times p' \times A$。在只考虑光路中的克服散粒噪声和非线性器件吸收的能量的情况下，该架构每次MAC运算的能

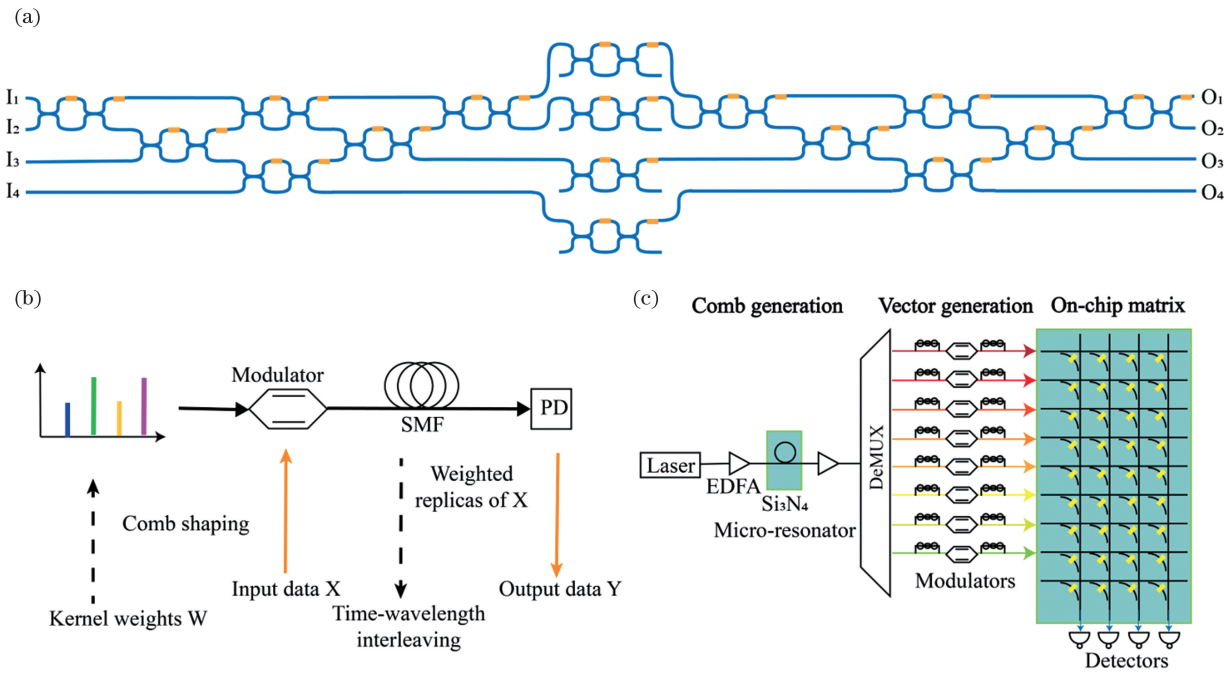


图5　三种典型的光电智能计算架构。(a)相干MZI网络[40]；(b)基于时间波长交织的光子加速器[62]；(c)基于PCM的集成光子张量核[61]

Fig. 5　Three typical optoelectronic intelligent computing architectures. (a) Coherent MZI mesh[40]; (b) photonic accelerator based on time-wavelength interleaving[62]; (c) integrated photonic tensor core based on PCM[61]

耗大约为30 fJ。

基于时间波长交织的卷积加速器如图5(b)所示。输入数据$X$被编码为时域电波形中时间符号的强度，波特率为$1/\tau$，其中$\tau$为符号周期。卷积核由长度为$R$($R$也是参与运算的波长数)的权重向量$\boldsymbol{W}$表示。根据文献[62]给出的光子卷积加速器的算力计算公式，其算力为$11.322\times10^{12}$ operation/s。在卷积神经网络应用中，卷积层的算力是$357\times10^{9}$ operation/s，全连接层的算力是$119.8\times10^{9}$ operation/s，合计是$476.8\times10^{9}$ operation/s。假设非线性器件的功耗是10 mW，卷积层和全连接层的最小光功率为72 mW，忽略光电转换的损耗，即$\eta=1$。光学频率梳的能耗为98 mW，其他电学设备的能耗总计200 mW。因此平均每次MAC运算的能耗为1.58 pJ。

基于PCM的集成光子张量核如图5(c)所示。在文献[61]展示的集成光子张量核中，$M=9\times4$，$V=4$，调制速率为14 GHz，因此可以计算得到其算力为$4.032\times10^{12}$ operation/s。在能效方面，根据文献[61]，具有16个WDM通道的$64\times64$阵列的能效为$5.1\times10^{12}$ operation·s$^{-1}$·W$^{-1}$。如果只考虑光路中散粒噪声的单位MAC能量下限，则每次MAC运算的能耗为17 fJ。

表2总结了不同架构的光电智能计算系统的算力和能效以及其评价方法其中$e_e$为光电计算架构的能效，$c_c$为光电计算架构的算力。由于集成光子张量核与MRR阵列的输入输出端口情况类似，因此其算力的评价方法也可以运用到MRR阵列上。但是考虑到MRR阵列本身的特点，MRR仅有1个输入向量，即$V=1$，可得MRR阵列的算力为

$$p=2\times M\times m_o。\tag{5}$$

表2　不同光电计算架构的算力和能效

Table 2　Computing capacity and energy efficiency of different optoelectronic computing architectures

| Performance | Coherent MZI mesh | Photonic accelerator based on time-wavelength interleaving | Integrated photonic tensor core based on PCM |
|---|---|---|---|
| Formula for computing capacity | $m\times2\times N^2\times10^{11}$ | $2R\times\frac{1}{\tau}$ | $2\times M\times m_o$ |
| Computing capacity | $6.4\times10^{12}$ operation/s | $11\times10^{12}$ operation/s | $4.302\times10^{12}$ operation/s |
| Formula for energy efficiency | $e_e=c_c/p_o$ | $e_e=c_c/p_o$ | $e_e=c_c/p_o$ |

根据前述的算力评价方法，可以分别从提高并行数、波特率和运算规模三个方面探索进一步提高算力的途径。图6系统总结了光电智能计算的算力和能耗的影响因素。

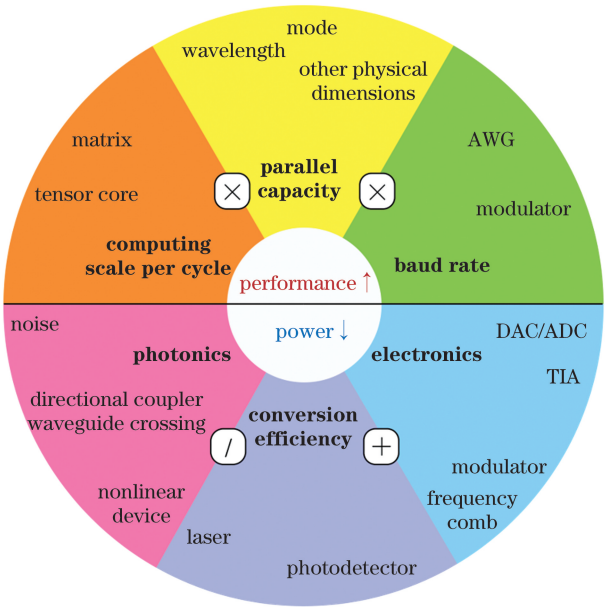


图6 算力和能耗的影响因素

Fig. 6 Factors influencing computing capacity and energy consumption

在提高并行数方面，首先可以通过增加参与运算的波长数量。我们往往采用光频梳提供多个波长的光源，在光频梳的各个梳齿线上编码信息，所以波长数的提高也往往依赖于光频梳性能的提升。提高并行数的另一方法是考虑其他维度的复用，例如采用多模MZI进行设计以复用模式这一维度。

在提升波特率方面，对于时间波长交织计算架构，可通过提升输入信号的速率，如任意波形发生器(AWG)的波特率。提升算力的另一重要方向是提升单次运算的规模。对于MZI、MRR、光子张量核等架构而言，单次运算规模就是片上的矩阵规模或者张量核的规模。然而，在提升矩阵规模时，往往会面临更严重的串扰问题，从而导致波特率提升方面的限制。

依据前文所述，在光子芯片中，降低能耗也可从三方面入手：降低光路中的能耗、提高光电/电光转换效率以及降低电学层的能耗。为了减少能量消耗，在光子芯片的设计中应尽量减少定向耦合器和交叉波导的使用，规避可能增大噪声的设计，同时使用更加节能高效的非线性光学器件。为减小噪声，可以在片上耦合激光器[83]，还可使用氮化硅[84]材料减小损耗。此外，也应提升具有光电/电光转换功能的激光器和探测器的外量子效率。

考虑消耗电能的器件，如调制器、光频梳、TIA和ADC/DAC等，由于电学器件的发展已趋于成熟，这部分功耗的降低主要取决于光学器件。对于调制器而言，使用与硅光子集成的铌酸锂和钛酸钡电光调制器能够以较低的功率实现高速的相位调制和低工作电压[85]。光电智能计算芯片通常需要大量幅度相位调控单元，通常采用金属热电极对光场进行调控，但一方面金属热电极需要能量维持工作状态，另一方面传统的金属热电极方案需要在金属层和波导层之间隔离一层厚的氧化物，因此大部分热量都被氧化层阻断，无法高效到达目标波导，导致热电极的加热效率和调节速率较低。针对传统热光相移器存在的问题，在降低能耗方面，可以使用纳米光电机系统(NOEMS)器件替代传统的热光相移器[86]，基于NOEMS的器件可以在没有静态能耗的情况下工作，机械位移只在切换到不同状态时需要能量。在光场高效调控方面，通过将石墨烯铺设在光子晶体慢光波导表面，可以实现加热效率极高的纳米热电极[87]，光信号开关时间仅需550 ns，与传统金属热电极相比，其调制速度提升了三个数量级。此外，侧面加热型的慢光增强的金属热电极[88]能够消除氧化硅隔离层的热传递损失，热调响应时间快至2 μs。根据非厄米系统理论，通过将热电极放置在离波导极近的地方，受益于热电极和波导之间的狭窄空间，热光相移器的带宽可以实现高达一个数量级的显著提升[89]。在存内计算方面，凭借PCM的非易失性，基于PCM的计算硬件不需要额外的能量来维持当前的配置，仅在状态切换时消耗能量。

## 4 结束语

光电智能计算充分融合了光学的多维复用、大带宽、低能耗等优势和电学的细粒度灵活控制特性，具有光算电控和软硬协同的特点，是一类面向人工智能计算的专用计算硬件架构，突破了传统冯·诺依曼架构的瓶颈。目前，光电智能计算领域正处于挑战与机遇并存的发展阶段。一方面，光电智能计算技术已经在光学神经网络和智能信号处理等诸多应用方面崭露头角，为实现低延迟、高带宽和低能耗智能计算开辟了道路。但另一方面，目前实验验证的光学神经网络大多是简单人工神经网络模型的演示系统，与成熟的电子神经网络相比，仅能够完成一些较为初级的深度学习任务，光电智能计算在大规模集成、在线训练、性能提升、产业化等方面仍然面临许

多挑战。光电智能计算是集成光子学、微电子学、人工智能和材料科学的交叉学科，其发展离不开多学科之间的密切交流与合作。先进的光电集成与制造技术为生产大规模和低成本的光电计算系统提供了可能，利用光电单片集成技术，将电子器件和光子器件集成在同一基底上，省去片上-片外的光电互连，构建片上光电混合协同计算架构，这将充分释放光电智能计算的潜力。此外，近年来基于新材料的光电子器件，如激光器、调制器和探测器，不断刷新各类器件的性能指标，将光电子器件与性能优异的新材料进行深度结合，有望进一步增强光电智能计算系统的性能。未来软硬件协同、光电融合一体的定制化设计将有力推动光电智能计算的发展，突破现阶段的难题和瓶颈，开启人工智能计算新时代。

## 参 考 文 献


[1] Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups[J]. IEEE Signal Processing Magazine, 2012, 29(6): 82-97.

[2] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[3] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[4] Grigorescu S, Trasnea B, Cocias T, et al. A survey of deep learning techniques for autonomous driving[J]. Journal of Field Robotics, 2020, 37(3): 362-386.

[5] Moore G E. Cramming more components onto integrated circuits[J]. Proceedings of the IEEE, 1998, 86(1): 82-85.

[6] Kim N S, Austin T, Baauw D, et al. Leakage current: Moore's law meets static power[J]. Computer, 2003, 36(12): 68-75.

[7] Lundstrom M. Moore's law forever?[J]. Science, 2003, 299(5604): 210-211.

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

[9] Benjamin B V, Gao P R, 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.

[10] Pei J, Deng L, Song S, et al. Towards artificial general intelligence with hybrid Tianjic chip architecture[J]. Nature, 2019, 572(7767): 106-111.

[11] 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.

[12] Graves A, Wayne G, Reynolds M, et al. Hybrid computing using a neural network with dynamic external memory[J]. Nature, 2016, 538(7626): 471-476.

[13] Davies M, Srinivasa N, Lin T H, et al. Loihi: a neuromorphic manycore processor with on-chip learning[J]. IEEE Micro, 2018, 38(1): 82-99.

[14] Huang C R, Sorger V J, Miscuglio M, et al. Prospects and applications of photonic neural networks[J]. Advances in Physics: X, 2022, 7(1): 1981155.

[15] Miller D A B. Device requirements for optical interconnects to silicon chips[J]. Proceedings of the IEEE, 2009, 97(7): 1166-1185.

[16] 陈宏伟, 于振明, 张天, 等. 光子神经网络发展与挑战[J]. 中国激光, 2020, 47(5): 0500004.
Chen H W, Yu Z M, Zhang T, et al. Advances and challenges of optical neural networks[J]. Chinese Journal of Lasers, 2020, 47(5): 0500004.

[17] Ahmed A H, Sharkia A, Casper B, et al. Silicon-photonics microring links for datacenters: challenges and opportunities[J]. IEEE Journal of Selected Topics in Quantum Electronics, 2016, 22(6): 194-203.

[18] 方璐, 戴琼海. 计算光场成像[J]. 光学学报, 2020, 40(1): 0111001.
Fang L, Dai Q H. Computational light field imaging[J]. Acta Optica Sinica, 2020, 40(1): 0111001.

[19] 左超, 冯世杰, 张翔宇, 等. 深度学习下的计算成像：现状、挑战与未来[J]. 光学学报, 2020, 40(1): 0111003.
Zuo C, Feng S J, Zhang X Y, et al. Deep learning based computational imaging: status, challenges, and future[J]. Acta Optica Sinica, 2020, 40(1): 0111003.

[20] 周治平, 许鹏飞, 董晓文. 硅基光电计算[J]. 中国激光, 2020, 47(6): 0600001.
Zhou Z P, Xu P F, Dong X W. Computing on silicon photonic platform[J]. Chinese Journal of Lasers, 2020, 47(6): 0600001.

[21] Li C, Zhang X, Li J W, et al. Correction to: the challenges of modern computing and new opportunities for optics[J]. PhotoniX, 2021, 2: 23.

[22] Cheng J W, Zhou H L, Dong J J. Photonic matrix computing: from fundamentals to applications[J]. Nanomaterials, 2021, 11(7): 1683.

[23] Zhou H L, Dong J J, Cheng J W, et al. Photonic matrix multiplication lights up photonic accelerator

and beyond[J]. Light: Science & Applications, 2022, 11: 30.

[24] Von B K. Lens design for optical Fourier transform systems[J]. Applied Optics, 1971, 10(12): 2739-2742.

[25] Athale R A, Collins W C. Optical matrix-matrix multiplier based on outer product decomposition[J]. Applied Optics, 1982, 21(12): 2089-2090.

[26] Labroille G, Denolle B, Jian P, et al. Efficient and mode selective spatial mode multiplexer based on multi-plane light conversion[J]. Optics Express, 2014, 22(13): 15599-15607.

[27] Lin X, Rivenson Y, Yardimci N T, et al. All-optical machine learning using diffractive deep neural networks[J]. Science, 2018, 361(6406): 1004-1008.

[28] Li J X, Mengu D, Luo Y, et al. Class-specific differential detection in diffractive optical neural networks improves inference accuracy[J]. Advanced Photonics, 2019, 1(4): 046001.

[29] Bernstein L, Sludds A, Hamerly R, et al. Freely scalable and reconfigurable optical hardware for deep learning[J]. Scientific Reports, 2021, 11: 3144.

[30] Pierangeli D, Marcucci G, Conti C. Large-scale photonic Ising machine by spatial light modulation[J]. Physical Review Letters, 2019, 122(21): 213902.

[31] Pierangeli D, Marcucci G, Conti C. Adiabatic evolution on a spatial-photonic Ising machine[J]. Optica, 2020, 7(11): 1535-1543.

[32] Zhou T, Lin X, Wu J, et al. Large-scale neuromorphic optoelectronic computing with a reconfigurable diffractive processing unit[J]. Nature Photonics, 2021, 15(5): 367-373.

[33] Tang R, Tanemura T, Nakano Y. Integrated reconfigurable unitary optical mode converter using MMI couplers[J]. IEEE Photonics Technology Letters, 2017, 29(12): 971-974.

[34] Tang R, Tanemura T, Ghosh S, et al. Reconfigurable all-optical on-chip MIMO three-mode demultiplexing based on multi-plane light conversion[J]. Optics Letters, 2018, 43(8): 1798-1801.

[35] Saygin M Y, Kondratyev I V, Dyakonov I V, et al. Robust architecture for programmable universal unitaries[J]. Physical Review Letters, 2020, 124(1): 010501.

[36] Tang R, Tanomura R, Tanemura T, et al. Ten-port unitary optical processor on a silicon photonic chip[J]. ACS Photonics, 2021, 8(7): 2074-2080.

[37] Reck M, Zeilinger A, Bernstein H J, et al. Experimental realization of any discrete unitary operator[J]. Physical Review Letters, 1994, 73(1): 58-61.

[38] Clements W R, Humphreys P C, Metcalf B J, et al. Optimal design for universal multiport interferometers[J]. Optica, 2016, 3(12): 1460-1465.

[39] Shokraneh F, Geoffroy-Gagnon S, Liboiron-Ladouceur O. The diamond mesh, a phase-error- and loss-tolerant field-programmable MZI-based optical processor for optical neural networks[J]. Optics Express, 2020, 28(16): 23495-23508.

[40] Shen Y, Harris N C, Skirlo S, et al. Deep learning with coherent nanophotonic circuits[J]. Nature Photonics, 2017, 11(7): 441-446.

[41] Xu S, Wang J, Shu H, et al. Optical coherent dot-product chip for sophisticated deep learning regression[J]. Light: Science & Applications, 2021, 10: 221.

[42] Zhou H L, Zhao Y H, Wei Y X, et al. All-in-one silicon photonic polarization processor[J]. Nanophotonics, 2019, 8(12): 2257-2267.

[43] Zhou H L, Zhao Y H, Xu G X, et al. Chip-scale optical matrix computation for PageRank algorithm[J]. IEEE Journal of Selected Topics in Quantum Electronics, 2020, 26(2): 8300910.

[44] Zhou H L, Zhao Y H, Wang X, et al. Self-configuring and reconfigurable silicon photonic signal processor[J]. ACS Photonics, 2020, 7(3): 792-799.

[45] Roques-Carmes C, Shen Y, Zanoci C, et al. Heuristic recurrent algorithms for photonic Ising machines[J]. Nature Communications, 2020, 11: 249.

[46] Prabhu M, Roques-Carmes C, Shen Y C, et al. Accelerating recurrent Ising machines in photonic integrated circuits[J]. Optica, 2020, 7(5): 551-558.

[47] Xu Q F, Soref R. Reconfigurable optical directed-logic circuits using microresonator-based optical switches[J]. Optics Express, 2011, 19(6): 5244-5259.

[48] Yang L, Ji R Q, Zhang L, et al. On-chip CMOS-compatible optical signal processor[J]. Optics Express, 2012, 20(12): 13560-13565.

[49] Tait A N, Nahmias M A, Shastri B J, et al. Broadcast and weight: an integrated network for scalable photonic spike processing[J]. Journal of Lightwave Technology, 2014, 32(21): 4029-4041.

[50] Tait A N, de Lima T F, Nahmias M A, et al. Multi-channel control for microring weight banks[J]. Optics Express, 2016, 24(8): 8895-8906.

[51] Tait A N, de Lima T F, Zhou E, et al. Neuromorphic photonic networks using silicon photonic weight banks[J]. Scientific Reports, 2017,

7: 7430.

[52] Ma P Y, Tait A N, de Lima T F, et al. Photonic principal component analysis using an on-chip microring weight bank[J]. Optics Express, 2019, 27(13): 18329-18342.

[53] Tait A N, de Lima T F, Nahmias M A, et al. Silicon photonic modulator neuron [J]. Physical Review Applied, 2019, 11(6): 064043.

[54] Huang C R, Bilodeau S, Ferreira de Lima T, et al. Demonstration of scalable microring weight bank control for large-scale photonic integrated circuits[J]. APL Photonics, 2020, 5(4): 040803.

[55] Ma P Y, Tait A N, de Lima T F, et al. Photonic independent component analysis using an on-chip microring weight bank[J]. Optics Express, 2020, 28(2): 1827-1844.

[56] Liu M, Zhao Y H, Wang X, et al. Widely tunable fractional-order photonic differentiator using a Mach-Zenhder interferometer coupled microring resonator [J]. Optics Express, 2017, 25(26): 33305-33314.

[57] Zhao Y H, Wang X, Gao D S, et al. On-chip programmable pulse processor employing cascaded MZI-MRR structure [J]. Frontiers of Optoelectronics, 2019, 12(2): 148-156.

[58] Cheng J W, Zhao Y H, Wei Y X, et al. On-chip photonic convolutional accelerator for image processing [C] // 26th Optoelectronics and Communications Conference, July 3-7, 2021, Hong Kong, China. Washington, D.C.: OPTICA, 2021: W4C.6.

[59] Feldmann J, Youngblood N, Wright C D, et al. All-optical spiking neurosynaptic networks with self-learning capabilities[J]. Nature, 2019, 569(7755): 208-214.

[60] Miscuglio M, Sorger V J. Photonic tensor cores for machine learning [J]. Applied Physics Reviews, 2020, 7(3): 031404.

[61] Feldmann J, Youngblood N, Karpov M, et al. Parallel convolutional processing using an integrated photonic tensor core[J]. Nature, 2021, 589(7840): 52-58.

[62] Xu X, Tan M, Corcoran B, et al. 11 TOPS photonic convolutional accelerator for optical neural networks [J]. Nature, 2021, 589(7840): 44-51.

[63] Bangari V, Marquez B A, Miller H, et al. Digital electronics and analog photonics for convolutional neural networks (DEAP-CNNs)[J]. IEEE Journal of Selected Topics in Quantum Electronics, 2020, 26(1): 7701213.

[64] Liao K, Chen Y, Yu Z C, et al. All-optical computing based on convolutional neural networks [J]. Opto-Electronic Advances, 2021, 4(11): 200060.

[65] Zang Y B, Chen M H, Yang S G, et al. Optoelectronic convolutional neural networks based on time-stretch method [J]. Science China Information Sciences, 2021, 64(2): 1-12.

[66] Guo X X, Xiang S Y, Zhang Y H, et al. Enhanced memory capacity of a neuromorphic reservoir computing system based on a VCSEL with double optical feedbacks [J]. Science China Information Sciences, 2020, 63(6): 160407.

[67] Rafayelyan M, Dong J, Tan Y Q, et al. Large-scale optical reservoir computing for spatiotemporal chaotic systems prediction[J]. Physical Review X, 2020, 10(4): 041037.

[68] Guo X X, Xiang S Y, Qu Y, et al. Enhanced prediction performance of a neuromorphic reservoir computing system using a semiconductor nanolaser with double phase conjugate feedbacks[J]. Journal of Lightwave Technology, 2021, 39(1): 129-135.

[69] Miscuglio M, Mehrabian A, Hu Z B, et al. All-optical nonlinear activation function for photonic neural networks [J]. Optical Materials Express, 2018, 8(12): 3851-3863.

[70] Zuo Y, Li B H, Zhao Y J, et al. All-optical neural network with nonlinear activation functions [J]. Optica, 2019, 6(9): 1132-1137.

[71] Wu B, Li H K, Tong W Y, et al. Low-threshold all-optical nonlinear activation function based on a Ge/Si hybrid structure in a microring resonator[J]. Optical Materials Express, 2022, 12(3): 970-980.

[72] Annoni A, Guglielmi E, Carminati M, et al. Unscrambling light-automatically undoing strong mixing between modes [J]. Light: Science & Applications, 2017, 6(12): e17110.

[73] Miller D A B. Analyzing and generating multimode optical fields using self-configuring networks [J]. Optica, 2020, 7(7): 794-801.

[74] Qu G Y, Yang W H, Song Q H, et al. Reprogrammable meta-hologram for optical encryption [J]. Nature Communications, 2020, 11: 5484.

[75] Goi E, Chen X, Zhang Q, et al. Nanoprinted high-neuron-density optical linear perceptrons performing near-infrared inference on a CMOS chip[J]. Light: Science & Applications, 2021, 10: 40.

[76] Khan M H, Shen H, Xuan Y, et al. Ultrabroad-bandwidth arbitrary radiofrequency waveform generation with a silicon photonic chip-based spectral shaper[J]. Nature Photonics, 2010, 4(2): 117-122.

[77] Hughes T W, Minkov M, Shi Y, et al. Training of photonic neural networks through *in situ* backpropagation and gradient measurement [J]. Optica, 2018, 5(7): 864-871.

[78] Zhou T K, Fang L, Yan T, et al. *In situ* optical backpropagation training of diffractive optical neural networks[J]. Photonics Research, 2020, 8(6): 940-953.

[79] Jouppi N P, Young C, Patil N, et al. In-datacenter performance analysis of a tensor processing unit[C]// ISCA'17: Proceedings of the 44th Annual International Symposium on Computer Architecture, June 24-28, 2017, Toronto, ON, Canada. New York: ACM Press, 2017: 1-12.

[80] NVIDIA. T4 tensor core datasheet[EB/OL]. [2021-11-30]. https: // www. nvidia. com/en-us/data-center/tesla-t4/.

[81] HUAWEI. A310 AI processor datasheet[EB/OL]. [2021-11-30]. https://e. huawei. com/cn/products/cloud-computing-dc/atlas/ascend-310.

[82] HUAWEI. A910 AI processor datasheet [EB/OL]. [2021-11-30]. https://e. huawei. com/cn/products/cloud-computing-dc/atlas/ascend-910.

[83] Mekis A, Gloeckner S, Masini G, et al. A grating-coupler-enabled CMOS photonics platform[J]. IEEE Journal of Selected Topics in Quantum Electronics, 2011, 17(3): 597-608.

[84] Baets R, Subramanian A Z, Clemmen S, et al. Silicon photonics: silicon nitride versus silicon-on-insulator[C]// 2016 Optical Fiber Communications Conference and Exhibition (OFC), March 20-24, 2016, Anaheim, CA, USA. New York: IEEE Press, 2016.

[85] Demkov A A, Bajaj C, Ekerdt J G, et al. Materials for emergent silicon-integrated optical computing[J]. Journal of Applied Physics, 2021, 130(7): 070907.

[86] Midolo L, Schliesser A, Fiore A. Nano-opto-electro-mechanical systems [J]. Nature Nanotechnology, 2018, 13(1): 11-18.

[87] Yan S Q, Zhu X L, Frandsen L H, et al. Slow-light-enhanced energy efficiency for graphene microheaters on silicon photonic crystal waveguides [J]. Nature Communications, 2017, 8: 14411.

[88] Yan S Q, Chen H, Gao S Q, et al. Efficient thermal tuning employing metallic microheater with slow-light effect [J]. IEEE Photonics Technology Letters, 2018, 30(12): 1151-1154.

[89] Wei Y X, Cheng J W, Wang Y L, et al. Fast-response silicon photonic microheater induced by parity-time symmetry breaking[EB/OL]. (2021-06-28)[2021-11-21]. https://arxiv.org/abs/2107.09444.

# Advances and Challenges of Optoelectronic Intelligent Computing

Cheng Junwei[1], Jiang Xueyi[1], Zhou Hailong[1], Dong Jianji[1,2*]

[1] *Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan* 430074, *Hubei, China*;

[2] *Optics Valley Laboratory, Wuhan* 430074, *Hubei, China*

## Abstract

**Significance** Artificial intelligence (AI) is one of the most active research fields at present. The AI models, represented by artificial neural networks, are computational models that mimic the neural synaptic networks in the brain and have been widely used in areas such as computer vision, speech recognition, and automatic driving. In the last decade, the AI technologies have experienced an explosive growth and the global computational volume has increased dramatically. The urgent need to process massive data in a fast and efficient way has placed an urgent demand on the computing hardware in terms of computing capacity and energy efficiency. Restricted by the inherent limits of electronic devices and the von Neumann architecture, traditional electronic computing has encountered a bottleneck in terms of speed and energy efficiency, which is difficult to break through. Optoelectronic intelligent computing uses photons instead of electrons to perform computation, hence optoelectronic intelligent computing can significantly improve computing speed and energy efficiency by overcoming the inherent limits of electrons. Compared with electronic computing, optoelectronic intelligent computing fully combines the unique advantages of multi-dimensional multiplexing, large bandwidth, and low energy consumption of optics and the fine-grained and flexible control of electronics, which is a more practical and competitive solution for accelerating AI computing.

Optoelectronic intelligent computing is particularly suitable for implementing large-scale neural networks containing a large number of neurons and synapses. Restricted by interconnection density and Joule heat, the processing speed of current neuromorphic electronic chips is basically limited to the MHz range, and the energy consumption per multiply accumulate (MAC) operation requires several picojoules. However, the neuromorphic

computing hardware built from basic photonic devices, such as Mach-Zehnder interferometer (MZI) mesh and microring resonator (MRR) array, requires only tens of femtojoules per MAC operation. This metric is two orders of magnitude smaller than that of the state-of-the-art complementary metal oxide semiconductor (CMOS) computing hardware, indicating that optical neural networks are far superior to electronic neural networks in terms of energy efficiency while achieving ultra-high-speed computing. As a result, optoelectronic intelligent computing naturally has significant advantages in the application scenarios such as automatic driving and drones which require large bandwidth and high real-time performance, as well as in the highly concurrent, high-throughput, computationally intensive supercomputing platforms and data centers. In fact, the optical interconnect technique has already been widely used in data centers and significantly reduces its time and energy consumption for large-scale interconnects.

Generally speaking, AI algorithms can be divided into two parts: training and inference. Since photons are difficult to be stored and the state of photons cannot be directly obtained, most of the current optoelectronic intelligent computing systems use the method of "training in the electronic domain and inference in the optical domain" to implement the AI algorithms. In other words, the simulation model of the neural network is first trained on the electronic computer, and then the parameters of the trained model are loaded onto the photonic chip for inference. However, this offline training method obviously has the difference between the the numerical simulation model and the actual physical model, and more importantly, if the neural network implementation of the photonic chip always needs to rely on the training in the electronic domain, then the performance advantages of the photonic chip over the microelectronic chip including low latency and high energy efficiency cannot be fully exploited. Therefore, developing hardware-friendly online training algorithms for optoelectronic intelligent computing is a key challenge.

**Progress** Here, a comprehensive review of the research progress and challenges in optoelectronic intelligent computing is presented. There are three mainstream types of optical matrix-vector multiplication (MVM), which are plane light conversion (PLC)-based matrix computing, MZI-based matrix computing, and wavelength division multiplexing (WDM)-based matrix computing (Fig. 1). The applications of optoelectronic intelligent computing mainly consist of optical signal processing and optical neural networks (Fig. 2). Hardware-friendly online training algorithms for optoelectronic intelligent computing mainly include online training of optical neural networks through the back propagation (BP) algorithm (Fig. 3) and online training of optoelectronic intelligent computing chips through the stochastic gradient descent (SGD) algorithm (Fig. 4). Computing capacity and energy efficiency are important metrics to evaluate the performance of optoelectronic intelligent computing. Table 1 shows the comparison of computing capacity and energy efficiency of various microelectronic chips and optoelectronic chips. As for three typical optoelectronic intelligent computing architectures (Fig. 5), their computing capacity and energy efficiency are summarized (Table 2). Finally, according to the aforementioned evaluation methods, the ways to further improve computing capacity and reduce energy consumption can be explored in terms of improving parallelism, baud rate, operation scale, and optical-electrical/electrical-optical conversion efficiency and reducing energy consumption in the optical and electrical layers (Fig. 6).

**Conclusions and Prospects** Optoelectronic intelligent computing fully combines the advantages of multi-dimensional multiplexing, large bandwidth, and low energy consumption of optics and the fine-grained and flexible control of electronics. It is a category of the dedicated computing hardware architectures for AI computing, which breaks the bottleneck of traditional von Neumann architectures. Here, the research progress of optoelectronic intelligent computing is reviewed, the challenges in online training algorithms as well as computing capacity and energy efficiency improvement of the current mainstream computing architectures for optical signal processing and optical neural networks are discussed, and the perspectives are presented. Advanced optoelectronic integration and fabrication techniques enable the production of large-scale and low-cost optoelectronic computing chips. Through optoelectronic monolithic integration, electronic and photonic devices can be integrated on the same substrate, which eliminates on-chip and off-chip optoelectronic interconnections and builds on-chip optoelectronic hybrid computing architectures. In addition, the performance of optoelectronic intelligent computing can be further improved by in-depth combination between novel materials with excellent performance and customized design with hardware-software synergy.

**Key words** optics in computing; optoelectronic intelligent computing; artificial intelligence; computing acceleration; optical signal processing; optical neural networks