·特邀综述·

光子神经网络发展与挑战

陈宏伟[1,2*]，于振明[3]，张天[3]，臧裕斌[1,2]，淡一航[3]，徐坤[3]

[1]清华大学电子工程系，北京 100084；

[2]北京信息科学与技术国家研究中心，北京 100084；

[3]北京邮电大学信息光子学与光通信国家重点实验室，北京 100876

摘要 近年来，以神经网络为代表的人工智能技术正向着高速低功耗的方向快速发展。然而，受限于电子器件的固有极限，传统电子神经网络难以进一步提高功率效率与计算速度。而光子神经网络能够有机地将光电子技术与神经网络模型相结合，提供了突破这一瓶颈的有效手段。为了更好地了解光子神经网络的发展历程，把握当前光子神经网络的研究热点以及展望未来光子神经网络的发展方向，本文对光子前馈、循环以及脉冲神经网络的研究现状进行梳理，以阐释光子神经网络在实时训练、非线性运算、规模化和实用化方面面临的挑战及未来发展的趋势。

关键词 光计算；神经网络；光电子技术；人工智能

中图分类号 O436 **文献标志码** A **doi**：10.3788/CJL202047.0500004

Advances and Challenges of Optical Neural Networks

Chen Hongwei[1,2*], Yu Zhenming[3], Zhang Tian[3], Zang Yubin[1,2], Dan Yihang[3], Xu Kun[3]

[1]*Department of Electronic Engineering, Tsinghua University, Beijing* 100084, *China*;

[2]*Beijing National Research Center for Information Science and Technology (BNRist), Beijing* 100084, *China*;

[3]*State Key Laboratory of Information Photonics and Optical Communications, Beijing University of Posts and Telecommunications, Beijing* 100876, *China*

Abstract Neural networks, as one of the most representative techniques in artificial intelligence, have been in rapid development towards high computational speed and low power cost. Due to intrinsic limitations brought by electronic devices, it can be hard for electronic implemented neural networks to further improve these two performances. Optical neural networks can combine both optoelectronic technique and neural network model to provide ways to break the bottleneck. In order to have a brighter view on the history, frontiers and future of optical neural networks, optical neural networks of feed-forward, recurrent and spiking models are illustrated in this paper. Challenges and future trends of optical neural networks on *in situ* training, nonlinear computing, expanding scale and applications will thus be revealed.

Key words optics in computing; neural networks; optoelectronic technology; artificial intelligence

OCIS codes 200.4700; 230.2090; 250.4390

1 引 言

人工智能，作为当今信息科学中最为活跃的领域之一，在机器视觉[1]、自动驾驶[2]、目标跟踪[3]及自然语言处理等领域[4-5]有着重要应用。人工神经网络，作为人工智能最重要的模型之一，因具有良好的泛化能力和鲁棒性而被广泛应用于各类场景之中。人工神经网络通过模仿神经系统结构，建立神经网络中各层神经元之间的连接。集成电路(IC)芯片是当今主流神经网络模型训练和测试的硬件载体，传统的神经网络可以在 CPU、GPU、现场可编程门阵列(FPGA)以及专用集成电路(ASIC)上运行。然而，无论采用何种 IC 芯片作为载体，其所采用的冯·诺伊曼结构都会将程序空间与数据空间分离，从而导致存储器与计算单元之间产生大量潮汐性数据荷载[6]。频繁的潮汐

收稿日期：2019-11-26；**修回日期**：2019-12-13；**录用日期**：2019-12-24

基金项目：北京市科技计划(Z181100008918011)、国家自然科学基金(61771284,61901045)

* **E-mail**：chenhw@tsinghua.edu.cn

性数据读写使得计算速率下降的同时增加了单次计算的功耗。当前，研究人员主要通过采用提高集成度和存内计算[6]这两种方案来提高运算效率，但这两种方案也面临很大挑战：一方面，通过不断缩小晶体管尺寸来提高运算速度的方法不可持续，不断缩小的晶体管尺寸将会带来逐渐显著的量子效应，使得进一步提升晶体管的运算效率变得很困难[7]；另一方面，存内计算的方案将面临大规模改造现行神经网络架构的难题，从而使适用于存内计算的神经网络算法的可移植性和兼容性下降[6]。也有研究人员采用频域计算[8]的方案来提高运算效率，但同样无法根除潮汐性数据读写带来的问题。除此之外，越来越高的运算速度和越来越大的神经网络规模对集成电路的频率响应提出了更高的要求，从而加大了芯片的制造难度。

光(电)子技术，是采用光子作为信息传输和处理基本载体的技术。相比传统的电子技术，光(电)子技术因具有大带宽、低损耗以及高传输信息量等优势，已在通信[9-10]、成像[11]、雷达[12]以及信号处理[13-14]领域被广泛应用。将该技术与传统神经网络模型结合，能够发挥光(电)子技术特有的优势，有望突破传统电神经网络长延时、高功耗等技术瓶颈。首先，光子神经网络采用存算一体的结构，规避了电神经网络存在的潮汐性数据读写问题，从而在提高计算速度的同时能够有效降低计算时延；其次，光子神经网络连接链路损耗较低，能有效提升功率效率；并且相比于传统电器件，光器件具有更大的带宽和更短的响应时间，因此更适应神经网络的高速实时计算。此外，针对自动驾驶、图像处理这类前端为光传感的应用领域，光子神经网络能够在物理层直接处理信息，从而可以避免光电转换引入的延时、功耗、信噪比劣化等问题。

目前，光子神经网络技术的研究包含了前馈神经网络、循环神经网络(储备池计算)以及脉冲神经网络这三种典型结构。同时，光子神经网络也正向着可实时训练、规模化以及特殊应用领域等方向继续发展。本文旨在回顾光子神经网络的发展，阐述目前已取得的阶段性进展，并揭示未来的发展趋势和面临的挑战。

2　光子神经网络的研究现状

2.1　前馈神经网络

前馈神经网络的基本特征是神经网络各层之间的信息由输入层向输出层单向传递，该网络一般包含全连接神经网络与卷积神经网络这两类结构。光子前馈神经网络中的矩阵运算主要是基于三角分解算法，该算法由 Reck 等[15]于 1994 年提出。三角分解算法基于无源无耗器件传输矩阵的一元性，证明了分束器、移相器所搭建的三角形网络能够实现任意参数和规模的酉矩阵[15]。2016 年，Ribeiro 等[16]利用这一分解算法，基于马赫-曾德尔干涉仪(MZI)设计了一个可实现任意 4×4 酉矩阵的集成芯片。同年，Clements 等[17]在 Reck 等方案的基础上进行优化设计，提出了矩形分解方案。结合奇异值分解理论，即任意矩阵可以分解成两个酉矩阵与一个对角矩阵的乘积，可以利用马赫-曾德尔干涉仪(酉矩阵)和可变衰减器(对角矩阵)实现任意矩阵。2017 年，Shen 等[18]基于上述思路成功研制了世界上第一款光子干涉计算单元芯片，他们先用该芯片实现线性运算，然后结合电域仿真的非线性激活函数构建了全连接神经网络。

图 1(a)所示为集成光子干涉计算单元(OIU)示意图，该芯片的输入输出端口数均为 4[18]，芯片内部的红色结构用来实现酉矩阵，蓝色结构用来实现对角矩阵。使用该芯片两次能够实现一层全连接神经网络的线性运算。该全连接神经网络中的非线性运算通过计算机仿真饱和吸收体的传输特性曲线实现。Shen 等利用该集成光子干涉计算单元搭建了含有单个隐藏层的全连接神经网络模型，实现了四个元音的分类。使用时，待识别的元音信号首先被调制于光脉冲上，之后通过芯片获得中间结果(调制信号与酉矩阵、对角矩阵相乘)。将该中间结果转化为电信号处理后调制于光脉冲上再次通过芯片，便能实现一层神经网络中的线性运算。该结果最后通过计算机仿真的非线性激活函数处理后得到一层网络的输出。重复两次上述操作即可实现该全连接网络，其识别元音的准确率可达 76.7%，识别的混淆矩阵如图 1(b)所示。

根据文献分析[18]，利用该光子干涉计算单元搭建的光子全连接神经网络具有如下优势：1)该芯片利用光(电)子技术将每层权重矩阵直接映射至 MZI 中移相器的相位上，从而规避了传统电神经网络面临的潮汐性数据荷载问题；2)未来，借助于相变材料(PCM)，该芯片的计算速度仅受芯片尺寸、色散模块谱宽以及光电探测器(PD)性能的限制，从而使得该芯片更能适应高速低功耗的神经网络运算；3)该芯片的计算时延将显著低于传统电神经网络。

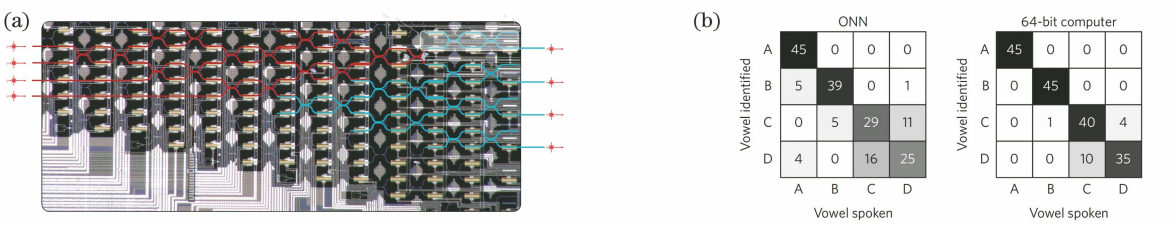

图 1　集成光子干涉计算单元结构和元音识别混淆矩阵[18]。(a)集成光子干涉计算单元结构;(b)元音识别混淆矩阵

Fig. 1　Structure of integrated optical interference unit (OIU) and confusion matrices of vowel recognition[18]. (a) Structure of OIU; (b) confusion matrices of vowel recognition by ONN and 64-bit computer

2018 年,Bagherian 等[19]在此光子干涉计算单元芯片的基础上提出利用该芯片通过时分复用的方式分段地计算图像卷积,从而构建更加复杂的卷积神经网络结构。构建出的卷积神经网络模型可用于彩色数字的识别。集成光子干涉计算单元输入端的延时线结构如图 2(a)所示,通过在芯片输入端口前附加延时线结构,实现了不同位置的卷积核与对应特征图的相乘累加计算。图 2(b)为使用芯片时的时序与逻辑步骤。首先,将输入图片按 RGB 三色切为垛叠的三片,并根据卷积核的大小对图片进行分块;然后,针对分割后的每一块子图,将其与同等大小的卷积核一并拉长成为向量并输入至光干涉芯片中;接着,将编码后的向量注入已配置好的光干涉模块中实现卷积运算;最后,将结果进行非线性处理,处理后就可以得到输出图样中的一个像素。针对不同的子图,重复上述操作便能得到一层卷积神经网络输出。这一工作所采用的时分复用方式能够充分利用芯片来实现图片卷积,将原来仅能进行四元音分类的简单全连接神经网络推广到了能够实现彩色图像识别的卷积神经网络,在保持系统低功耗、高运算速度的优势下,进一步提升了网络的复杂度。

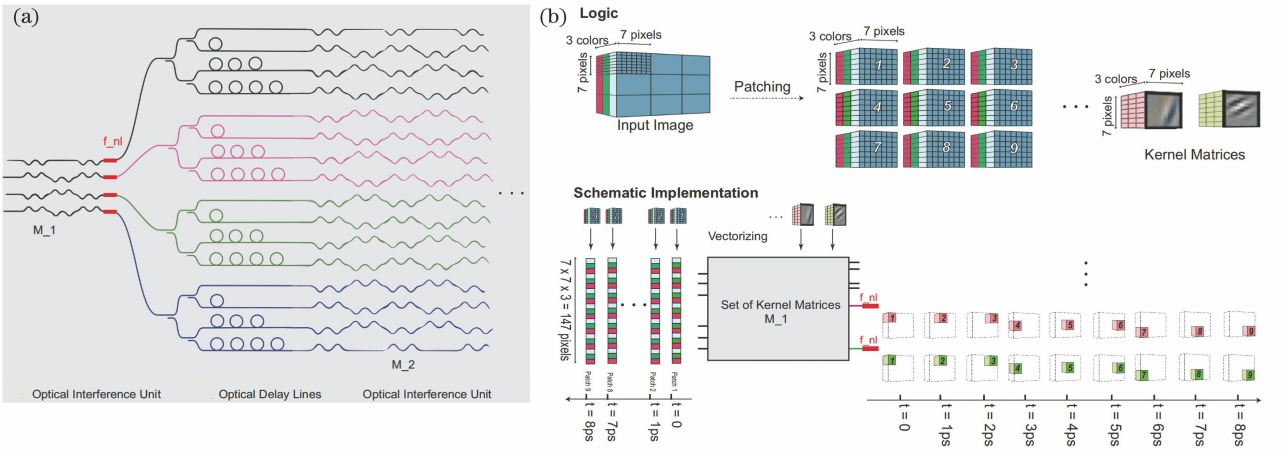

图 2　集成光子干涉运算单元芯片前的延时线结构和光卷积神经网络实现的时序逻辑图[19]。(a)延时线结构;(b)时序逻辑图

Fig. 2　Structure of delay line before OIU and logic and timing chart of the optical convolutional neural network[19]. (a) Structure of delay line; (b) logic and timing chart

2.2　循环神经网络(存储池计算)

与前馈神经网络不同,循环神经网络(又被称为存储池计算)各层之间的信息除了单向传递外,还存在节点间的连接与后向反馈连接[20-22]。如图 3(a)所示,循环神经网络主要由输入层、中间层(又称为存储池)以及输出层组成。除了各层之间的前馈连接 W_{in}、W_{out}之外,还存在存储池内部节点之间的互连 W_{res}以及输出层至存储池之间的反馈 W_{back}。训练时只训练 W_{out}就能使整个网络实现相应的功能。

使用光电子器件构建循环神经网络时,存在两种方案。第一种是采用如图 3(a)所示的并行结构,即循环神经网络的每一个节点都用一个光电子器件搭建。这种方案的优点是直观性较好,并且得益于并行计算,计算速度较快。2011 年,Vandoorne 等[23]利用半导体光放大器(SOA)的增益饱和效应和动态弛豫特性来分别构建循环神经网络存储池的非线性连接及反馈,通过仿真,该循环神经网络模型能够实现语音识别。2016 年,Bueno 等[24]使用空间

光学器件搭建的循环神经网络也采用了并行结构。但并行结构存在鲁棒性较差、规模不易扩展以及成本较高等问题。为了解决这一问题，第二种方案，即如图 3(b)所示的串行结构被提出。在该结构中，存储池内原有的众多节点采用单一非线性节点以延时的方式实现。

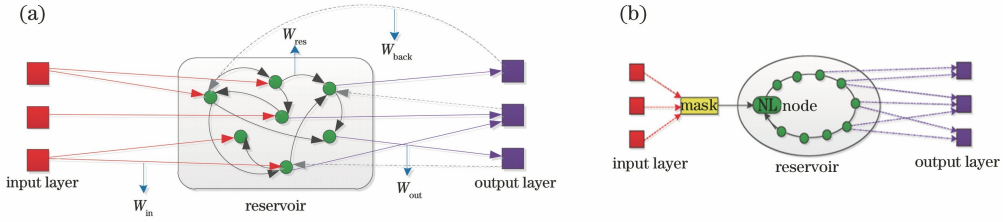

图 3　循环神经网络的结构[20]。(a)并行结构；(b)串行结构

Fig. 3　Structures of recurrent neural networks[20]. (a) Parallel structure; (b) serial structure

2012 年，Paquot 等[25]利用光电混合系统率先实现了光电混合串行循环神经网络，其结构如图 4(a)所示，信号从任意波形发生器注入，通过放大器和调制器调制于光上。中间的存储池由可变光衰减器、延时环、反馈光电二极管、混频器、放大器以及马赫-曾德尔调制器构成。光电二极管将系统的输出转化为电信号并读出。通过对输出权重进行训练与控制，系统可以实现方波与正弦波的识别。如图 4(b)所示，已量化的正弦波和方波串行输入到系统中，通过循环神经网络的处理，系统将串行地输出如图 4(b)下图所示的分类结果，其中 0 代表正弦波信号，1 代表方波信号。此外，该系统还可以实现非线性信道均衡以及数字语音识别等功能。

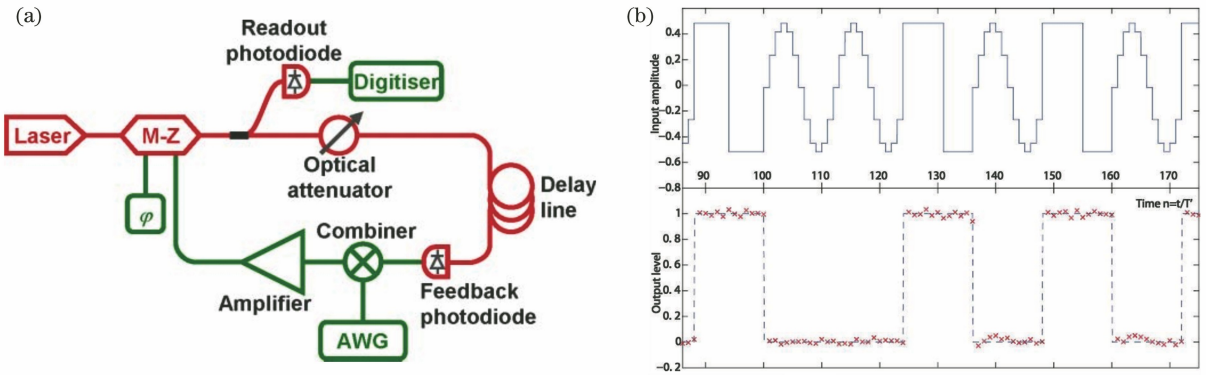

图 4　光电混合循环神经网络结构以及方波和正弦信号的分类结果[25]。(a)光电混合循环神经网络结构；(b)分类结果

Fig. 4　Structure of optoelectronic recurrent neural network and classifying result of rectangular and sinusoid signals[25]. (a) Structure of optoelectronic recurrent neural network; (b) classifying result

同年，Duport 等[26]利用超辐射发光二极管、调制器、任意波形发生器、光带通滤波器、光可调衰减器、分束器、隔离器、半导体光放大器、延时线以及光电二极管成功搭建了全光串行循环神经网络。除了直接采用延时线获得延时功能外，近年来，基于微环、多模干涉等器件在实现串行神经网络中的延时功能方面也取得了一定进展[27]。同时，此类光子循环神经网络也尝试采用多级或更加复杂的时分复用方式，以进一步提高运算速度[27]。

2.3　脉冲神经网络

脉冲神经网络(SNNs)又被称为第三代人工神经网络[28]。与前馈神经网络和循环神经网络相比，脉冲神经网络的模拟神经元更接近生物学神经元模型，除此之外，脉冲神经网络中的神经元并非在每一次迭代传播中都被激活，而是只有当其膜电位达到阈值时才被激活。当一个神经元被激活时，它会产生一个信号并将其传递给其它神经元，提高或降低其级联的神经元膜电位。在脉冲神经网络中，神经元的当前激活水平通常被建模成某种微分方程，其当前激活水平会在刺激脉冲到来后升高并持续一段时间，然后逐渐衰退。脉冲神经网络增强了处理时空数据的能力：一方面，脉冲神经网络结构中的神经元仅与附近的神经元连接，分别处理输入块，从而增强了空间信息的处理能力；另一方面，由于训练依赖脉冲时间间隔信息，因此二进制编码中丢失的信息可以在脉冲的

时间信息中重新获取，从而增强了时间信息的处理能力。事实证明，脉冲神经元是比传统人工神经元更强大的计算单元，是未来的一大发展趋势。然而，由于脉冲神经网络的训练方法和硬件实现还存在较多困难，因此暂未获得广泛应用，脉冲神经网络的大部分研究还集中在理论研究和简易结构的验证阶段。但是，现在已有更多的研究人员投入到脉冲神经网络训练算法和硬件（光学）实现的研究中来。

2016 年，普林斯顿大学的 Prucnal 研究小组提出了基于可激活的石墨烯光纤激光器的脉冲处理系统[29]，其结构如图 5 所示。该系统主要由掺铒光纤（增益部分）和石墨烯饱和吸收体（损耗部分）构成，980 nm 激光器充当泵浦源，1480 nm 激光器携带脉冲刺激信号激发系统产生类 LIF（leaky integrate-and-fire）脉冲神经元的响应。

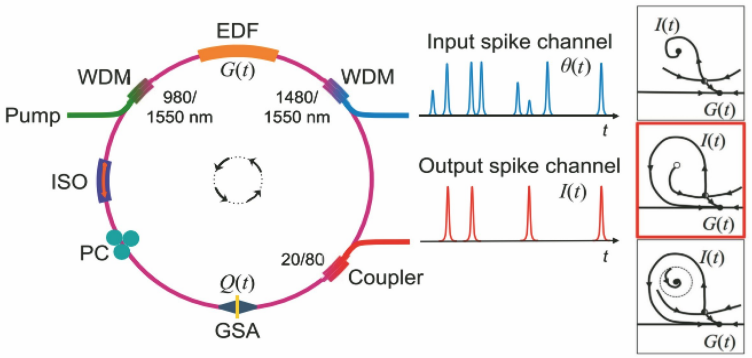

图 5　可激活的石墨烯光纤激光器系统[29]

Fig. 5　Graphene excitable fiber laser[29]

2018 年，该研究小组提出了基于分布式反馈（DFB）激光器结构的神经拟态光子集成电路[30]，集成的光子神经元如图 6（a）所示，每个 DFB 单元对刺激的响应如图 6（b）所示。

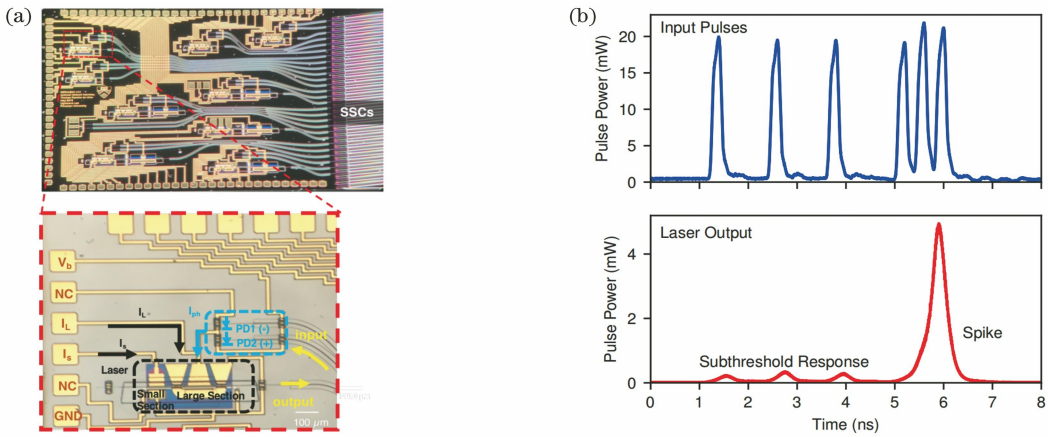

图 6　集成光子神经元和神经元对刺激的响应[30]。（a）集成光子神经元；（b）神经元对刺激的响应

Fig. 6　Integrated photonic neuron and excitable response[30]. (a) Integrated photonic neuron; (b) excitable response

此外，该研究小组探讨了构建可编程可级联光子神经网络的可行方案，内容包括：级联性的研究、Broadcast-and-weight（B & W）网络原型[31]、相干光方案。其中，B & W 网络原型是该研究小组提出的一种可支持大规模光子脉冲神经元并行互联的网络架构。

脉冲神经网络的训练算法主要以监督学习算法为主。监督学习算法包括：1）基于梯度下降的 SpikeProp[32] 算法，该算法适用于多层前馈脉冲神经网络；2）Tempotron 算法[33]，该算法根据输出层神经元输出的脉冲序列和期望序列的差别调整权重，训练的目的是使实际输出神经元膜电位的变化符合期望值，但该算法中的神经元仅有 0 和 1 两种输出类型，无法扩展到多层网络结构；3）基于突触可塑性（STDP）的算法，如 Hebbian 学习算法[34]；4）远程监督学习算法，如 ReSuMe 算法[35]；5）基于脉冲序列卷积的监督学习算法，如 SPAN 算法[36] 和 PSD 算法[37]。

3　光子神经网络的发展趋势

光子神经网络作为人工智能与光(电)子技术的交叉领域，能够兼具两者的长处。无论是前馈神经网络、循环神经网络还是脉冲神经网络，采用光电子器件构建均能在一定程度上提升运算速度、降低功耗。然而，受限于目前光电子器件的制作精度及实现难度，大部分光子神经网络仍然存在难以实时训练、非线性运算实现困难、神经网络规模与应用受限等问题。这些问题既严重制约着光子神经网络的发展，也限制了其产业化。

3.1　实时训练算法

光子神经网络，特别是前馈网络的训练问题是扩展光子神经网络应用的制约因素。由于光子本身无法像电子一样存储，因此无法直接对光子的状态进行记录，故而在电神经网络训练中应用广泛的后向传播算法难以移植于光子神经网络的训练上。针对这一问题，Hughes 等[38]于 2018 年首次提出了片上训练算法，其算法流程如图 7 所示。应用该算法，通过记录光场分布以及移相器的相位分布能够得到向收敛方向下降的梯度值，进而计算下一轮迭代中芯片移相器的相位配置，从而使得芯片整体性能能够逐步收敛到一个较好的结果。

Hughes 等通过仿真的方法在片上训练了一个具有两个光干涉计算单元(OIU)的神经网络来实现异或逻辑，以验证算法有效性。

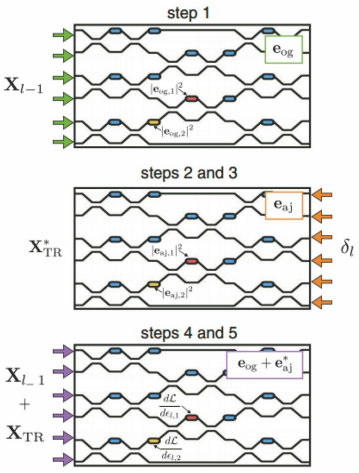

图 7　光神经网络芯片的片上训练算法[38]

Fig. 7　*In situ* training algorithm of integrated photonic neural networks[38]

未来，该算法能够进一步扩展并移植至更大规模的光子神经网络芯片中，通过片上训练使芯片逐渐收敛逼近最佳配置，从而完成特定功能。

当前，光子神经网络(ONNs)的训练方法主要还是后向传播(BP)算法和随机梯度下降(SGD)算法[38]。然而，BP 和 SGD 训练策略在集成光学芯片中很难实现，因此大部分光子神经网络需通过数字计算机进行模型预训练获取权值[18]。显然，这种训练方法不仅效率低下，而且由于模型表示的准确性受限，在速度和功耗方面失去了优势。Zhang 等[39]于 2019 年提出了一种基于神经进化(neuroevolution)策略的有效训练算法，即分别使用遗传算法(GA)和粒子群优化算法(PSO)来训练光子神经网络中的超参数并优化连接权重。训练后的光子神经网络被用来完成分类任务以进行性能评估，计算结果显示，其准确率和稳定性足以与传统的学习算法相竞争。为了验证该算法的有效性，Zhang 等[39]通过仿真分别实现了光神经网络在 Iris 数据集、Wine 数据集上的在线训练。Zhang 等[39]还用光子神经网络实现了此前 通常利用电神经网络[40-41]实现的通信信号调制格式的分类。上述三类数据集的分类效果随着训练次数的变化如图 8(a)与图 8(b)所示。

3.2　非线性运算

非线性运算已成为光子神经网络发展的另一个瓶颈，其原因在于光电子器件相比电子器件实现非线性函数更加困难，并且所实现的非线性函数存在很多非理想特性。然而，神经网络中的非线性函数可以加快网络的收敛速度，提升识别准确率，已成为神经网络中不可缺少的组成部分。当前，在光子神经网络中非线性运算的实现手段主要是利用饱和吸收体或电路仿真，但这些方法有的难于精准控制，有的则需要将光信号通过光电二极管转化为电信号，如此便会降低计算速度。

当前，研究较多的全光非线性运算元件是饱和吸收体。以饱和吸收体作为非线性运算元件的原理是将饱和吸收体的传输特性曲线作为神经网络中的激活函数。随着入射脉冲峰值光强增加，饱和吸收体的吸收系数逐渐减小，使得光透过率增大。1967 年，Selden[42]通过仿真得到了饱和吸收体的传输特性曲线。利用该曲线可以近似模拟神经网络中运用广泛的非线性函数。在 Seldon 模型被提出之后，研究人员进一步研究了饱和吸收体的非线性特性，并尝试将其应用在前馈及循环神经网络的非线性部分[43-44]。

2019 年，Williamson 等提出了一种光电混合可控的非线性运算模块[45]，其基本结构如图 9(a)所示。该模块由定向耦合器、延时线、MZI、光电二极

管、放大器、偏置电压源以及电传输模块组成。输入的携带有线性计算结果的光信号通过定向耦合器分成两路，一路通过延时线传递至 MZI 的输入臂，另一路通过光电二极管转化成电信号在电域上经过相应处理后与偏置电压一起控制 MZI 上移相器的相位，最后干涉结果从 MZI 的一臂输出。如图 9(b)所示，通过调节移相器的电压可以改变非线性函数，从而实现非线性模块的可重构性。

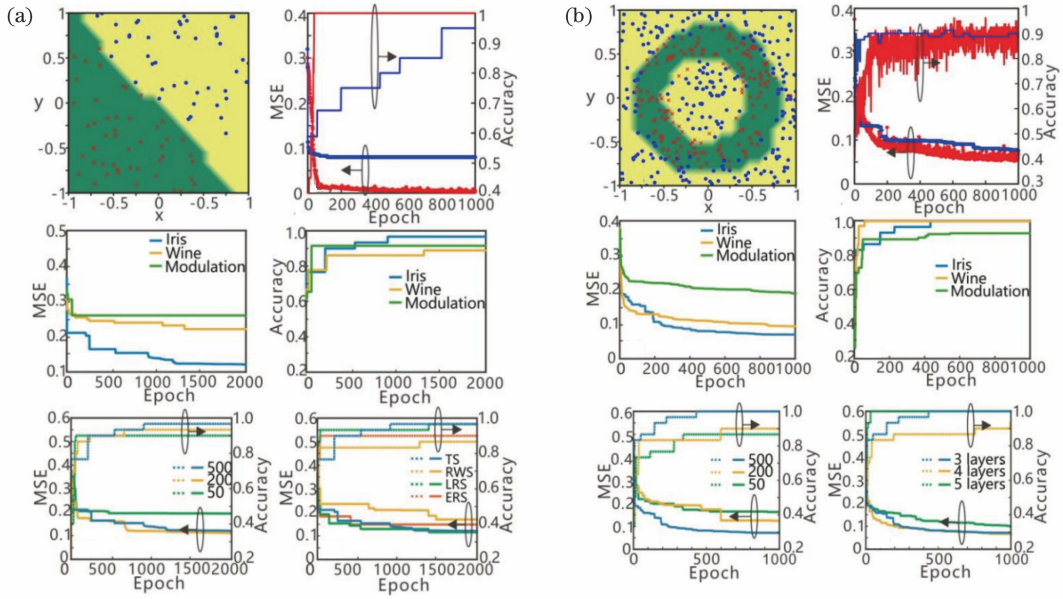

图 8　采用不同算法训练光神经网络的分类结果图[39]。(a) GA 算法；(b) PSO 算法

Fig. 8　Training results of optical neural network by adopting different algorithms[39]. (a) GA algorithm; (b) PSO algorithm

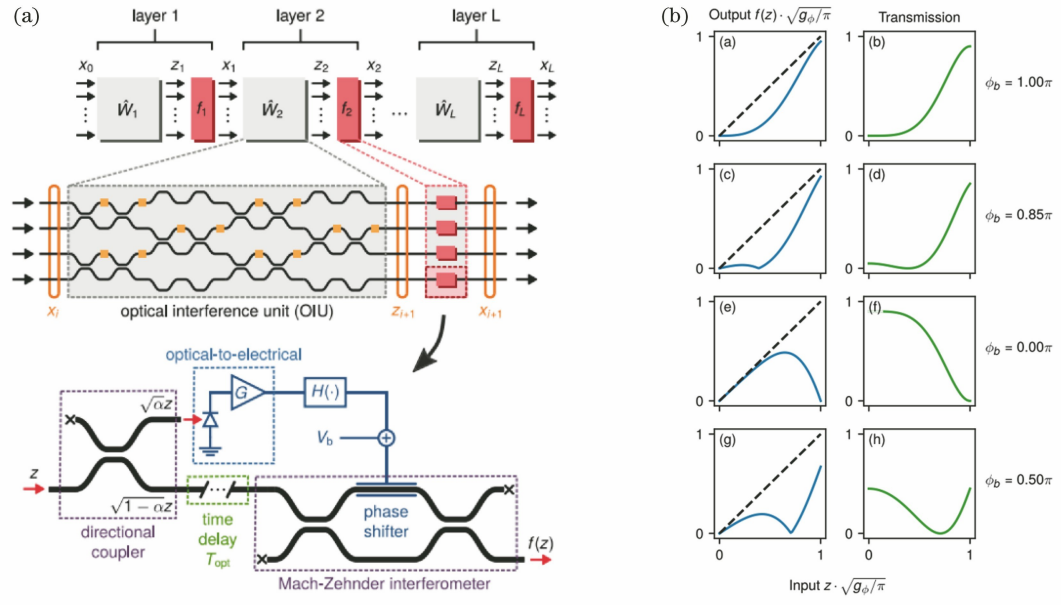

图 9　光电混合非线性模块结构以及通过调谐相位实现不同的非线性函数[45]。(a)非线性模块结构；(b)非线性函数

Fig. 9　Structure of optoelectronic nonlinear module and implementing different nonlinear functions by tunning phase of MZI[45]. (a) Structure of nonlinear module; (b) implementing different nonlinear functions

除了上述两种方案之外，2019 年，Feldmann 等[46]提出了光控相变存储器(PCM)方案，并采用该方案实现了光子神经元的非线性变换。光控 PCM 是一类工作状态在晶体状态和非晶状态之间的材料，其工作状态受输入光功率的控制。如图 10 所示，当输入光功率低于阈值功率时，PCM 处于晶体状态，大量的光功率被吸收；当输入光功率高于阈值功率时，PCM 处于非晶形状态，大部分光功率可以

通过。因此，将该材料集成于光传输介质中，就可以根据输入光功率改变材料的光通透性，从而实现光子神经网络非线性激活函数的功能。

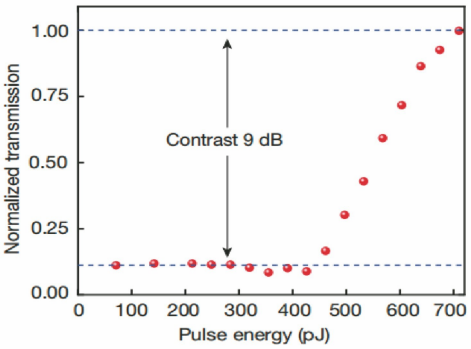

图 10　光控 PCM 的归一化传输系数曲线[46]

Fig. 10　Normalized transmission curve of optical controlled PCM[46]

二维石墨烯材料也有望用来实现光学激活函数[47-50]。与传统的饱和吸收体相比，石墨烯具有阈值低、易激发、非线性效应丰富等特点。目前，针对光学非线性材料与器件的研究正在进行[51-52]，研究目标集中在低阈值、可重构、易集成以及快速响应这四个方面。

3.3　光子神经网络的规模化

光子神经网络的规模化是另一个较为难解决的问题。一方面，大规模神经网络有利于实现更加复杂的功能，另一方面，光子器件的不稳定和难以精细调谐的特性又使得扩展神经网络的规模变得困难。为了解决这一问题，2018 年 Lin 等[53]提出了一种基于衍射的光子深度神经网络（D^2NN）结构，该结构不仅实现了手写数字和 Fashion 数据集的分类功能，还提升了空间成像分辨率。图 11(a)为基于衍射的光子深度神经网络结构图。该光子深度神经网络基于惠更斯原理实现线性运算，利用光衍射叠加原理实现相邻两层神经元之间的连接。使用时，相干光平行入射“相位调制”板。“相位调制”板上的每个像素块相当于神经元，处于不同位置的像素块通过改变厚度来改变光经过时的相位差，从而使得不同节点之间有不同的权值。经过这样一系列传输后，通过统计放置于特定位置的 PD 的接收结果就可以获得深度神经网络的输出。由于该“相位调制”板具有数以百计的像素点，因此采用该结构的神经网络的每层节点数可以扩充至几百。图 11(b)为该神经网络对空间成像分辨率的提升效果图。

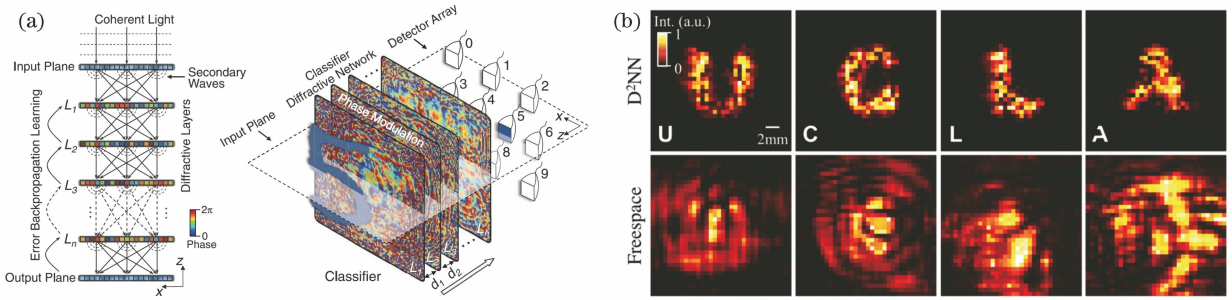

图 11　基于衍射的光子深度神经网络结构图及其对空间成像分辨率的提升效果[53]。(a)结构图；(b)提升效果

Fig. 11　Structure of D^2NN and resolution improvement of imaging[53]. (a) Structure D^2NN; (b) resolution improvement result

为扩大光子神经网络的维度，也可以用时间换取空间，通过串行的方式，在运算速度降低不大的情况下扩展神经网络的规模。清华大学某课题组提出了一种基于时域拉伸的串行光子神经网络（TS-NN）[54]，其基本结构如图 12 所示。从激光器输出的宽谱脉冲被大色散模块色散展宽并进行光谱整形，之后的调制器分别将每层的输入及配置调制到光脉冲上。在经过色散压缩模块以及光电二极管之后，携带有信息的光脉冲的能量被压缩积累，最终得到向量相乘、相加的计算结果。该方法通过并行变串行的方案实现了光电混合的全连接神经网络。采用该方案可以实现大规模的神经网络。理论上，单个展宽脉冲能够模拟的神经网络节点数取决于展宽光脉冲的宽度以及任意波形发生器的最大模拟带宽。

利用该结构，通过仿真验证了一个具有 3 层，每层分别含有 400、23、10 个神经元的神经网络，以实现手写数字识别的功能。图 13(a)～(c)分别显示了传统电神经网络结构、TS-NN 结构在无噪声情况下以及有噪声情况下的混淆矩阵，其识别准确率分别为 94%、89%与 88%[54]。

3.4　光神经网络的实用化

光子神经网络的另一个发展趋势是实用化。然而，受限于现有光子神经网络的规模与复杂度，光子神经网络的实用化还需要一段很长的发展时间。在

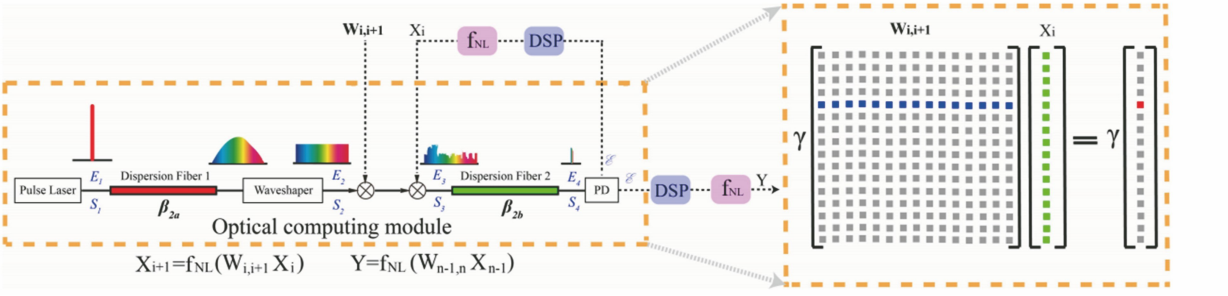

图 12　基于时域拉伸的全连接光电神经网络结构[54]

Fig. 12　Structure of TS-NN[54]

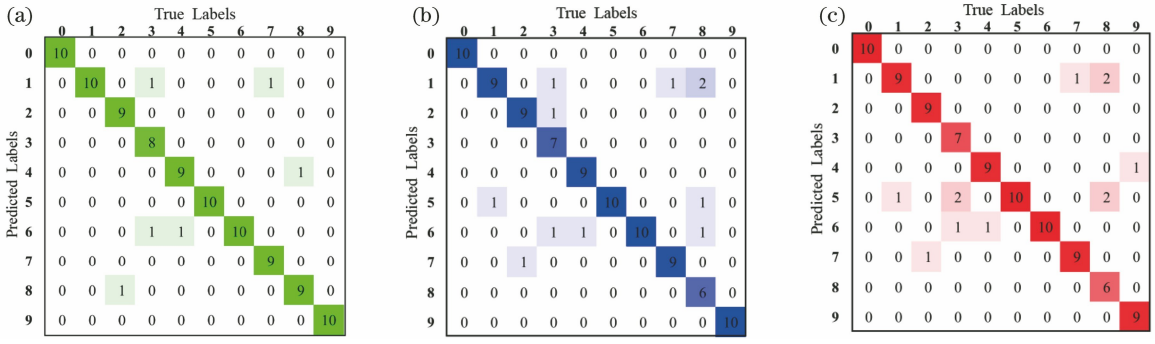

图 13　不同神经网络结构的识别混淆矩阵[54]。(a)传统神经网络；(b) TS-NN 网络在无噪声情况下；(c) TS-NN 结构在有噪声情况下

Fig. 13　Confusion matrices of different neural network structures[54]. (a) Conventional neural network structure; (b) TS-NN without noise; (c) TS-NN with noise

通信与数据处理领域，Paquot 等[25]于 2012 年构建的基于光纤系统的光电混合循环神经网络实现了对通信信道的均衡功能。此外，得益于循环神经网络在时间序列信号回归分析上的优势，光存储池计算目前已在金融序列预测、序贯信号处理上获得了初步应用[24-26]。在光通信领域，2019 年，Yu 等[55]利用光电混合网络实现了二值光相干接收机，实现了对发送端调制信号的恢复。其结构如图 14(a)所示，正交相移键控(QPSK)调制的光信号从输入层输入后，先通过二值全光神经网络进行处理，之后通过光电二极管和模数转换器变为电信号，最后利用电人工神经网络恢复原始调制信号。Yu 等针对单偏振和偏振复用系统分别提出了两种在光域上实现二值权重映射的结构，如图 14(b)和图 14(c)所示，通过移相器、混频器、平衡光电二极管实现二值神经网络的线性计算单元。利用平衡光电二极管和 1 位垂直分辨率的模数转换器(ADC)实现二值神经网络的激活函数，经过平衡光电二极管和 ADC 的电信号再通过电神经网络继续处理，最终可以实现发射信号的恢复。

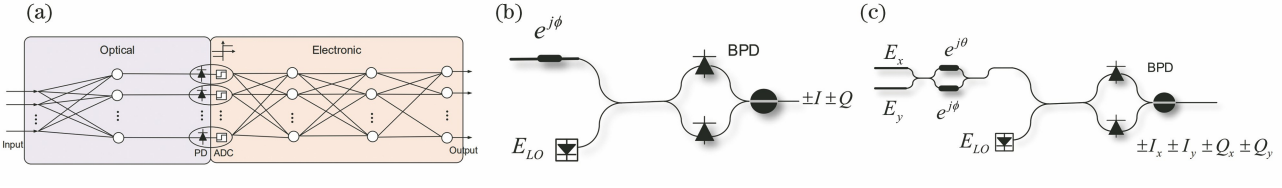

图 14　光电混合二值神经网络架构以及映射结构[55]。(a)光电混合二值神经网络架构；(b)单偏振系统的映射结构；(c)偏振复用系统的映射结构

Fig. 14　Structures of optoelectronic binarized neural network and weights mapping[55]. (a) Structure of the optoelectronic binarized neural network; (b) mapping structure of binarized weights in single polarization system; (c) mapping structure of binarized weights in polarization multiplexing system

通过仿真可知，该光电混合二值神经网络分别实现了 50 Gb/s 单偏振 SP-QPSK 信号接收机以及 100 Gb/s 偏振复用 QPSK 信号接收机。通过对比通信链路长度(背靠背，5 km 单模光纤)、激光器线

宽(0,100 Hz)以及实现方式(传统接收机,光电混合二值神经网络),得到了如图 15(a)和图 15(b)所示的误码率(BER)-光信噪比(OSNR)曲线。仿真结果表明该二值相干接收机可以用于相干光通信系统,光域的神经网络计算可以缓解电域的信号处理压力,降低整体光接收机的功耗,提升光接收机的信号处理速度。此外,该二值神经网络还能极大地降低对模数转换器量化位数的要求,使得仅用 1 bit 量化的模数转换器便能实现复杂调试格式信号的恢复,极大地降低了光接收机的成本。

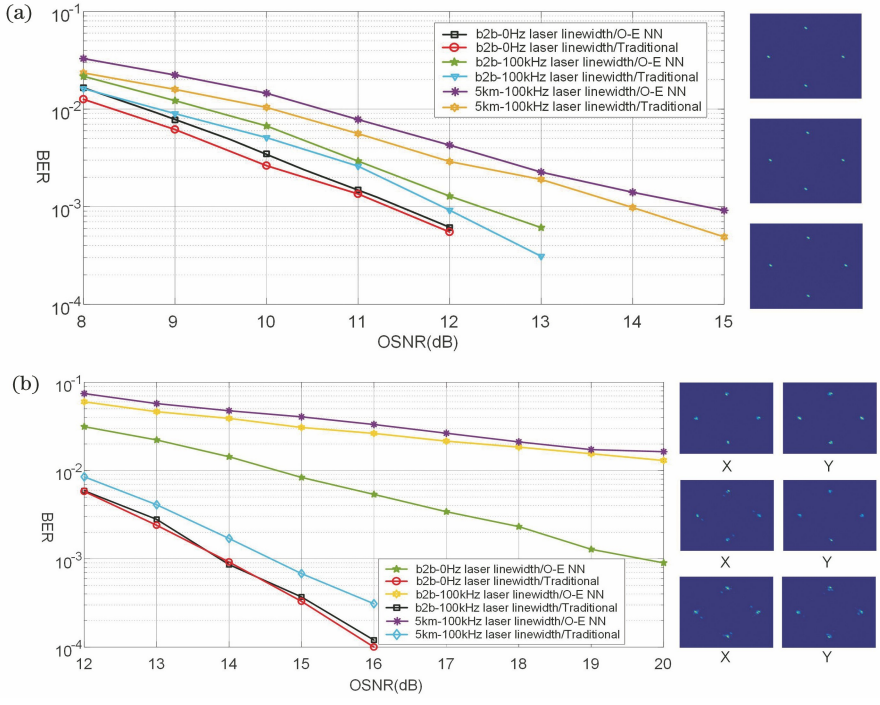

图 15　以不同方式实现的单偏振、偏振复用 QPSK 通信系统的误码率-光信噪比曲线[55]。(a)单偏振 QPSK 通信系统;(b)偏振复用 QPSK 通信系统

Fig. 15　OSNR-BER curves of single polarization and polarization multiplexing QPSK communication systems implemented by different methods[55]. (a) Single polarization QPSK communication system; (b) polarization multiplexing QPSK communication system

4　结束语

光子神经网络作为光(电)子技术与人工智能技术的交叉学科产物,能够将两者的优势结合起来,构建出高速低功耗的网络结构,突破传统电子神经网络的瓶颈。得益于光电子器件制造技术的成熟和发展,尤其是集成光电子技术的发展,光子神经网络技术在利用光电子器件构建前馈、循环与脉冲神经网络方面都取得了突破性进展。然而,相较于目前发展得较为成熟的电子神经网络,光子神经网络在可训练性、集成度、规模化、实用化等方面仍然具有广阔的提升空间。一方面,光电子器件性能的非理想性与低稳定性抑制了光子神经网络的可训练性、集成度与规模化,为构建功能更加复杂的神经网络模型提出了更为严苛的要求;另一方面,光子神经网络在应用领域也受到了上述效应的限制,难以在特定领域充分发挥优势。近年来,虽然众多解决方案被提出,但是如何从根源上突破光子神经网络的瓶颈还需要深入思考和研究。相信在不久的将来,光子神经网络一定能够克服这些难题,更好地发挥光电子技术与人工智能技术带来的高速低功耗优势,更好地构建绿色智能的世界。

参　考　文　献

[1] LeCun Y, Bengio Y. Convolutional networks for images, speech, and time series[J/OL]. [2019-11-20]. https://www.researchgate.net/publication/2453996_Convolutional_Networks_for_Images_Speech_and_Time-Series.

[2] Al-Qizwini M, Barjasteh I, Al-Qassab H, et al. Deep learning algorithm for autonomous driving using GoogLeNet[C]// 2017 IEEE Intelligent Vehicles Symposium (IV), June 11-14, 2017. Los Angeles, CA, USA. NewYork: IEEE, 2017: 89-96.

[3] Garg R, Vijay K B G, Carneiro G, et al. Unsupervised CNN for single view depth estimation:

geometry to the rescue[M]// Computer Vision-ECCV 2016. Cham: Springer International Publishing, 2016: 740-756.

[4] 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.

[5] Sercu T, Puhrsch C, Kingsbury B, et al. Very deep multilingual convolutional neural networks for LVCSR[C]// 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), March 20-25, 2016. Shanghai. [s.n.]: IEEE, 2016: 4955-4959.

[6] Zaharia M, Chowdhury M, Das T, et al. Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing[C/OL]// Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation. [2019-11-20]. http: // citeseerx. ist. psu. edu/viewdoc/download; jsessionid=E2EF85C42E9B7DB99EF31A33BAE02668? doi=10.1.1.225.856&rep=rep1&type=pdf.

[7] Waldrop M M. The chips are down for Moore's law [J]. Nature, 2016, 530(7589): 144-147.

[8] Highlander T, Rodriguez A. Very efficient training of convolutional neural networks using fast Fourier transform and overlap-and-add[C/OL]// Procedings of the British Machine Vision Conference 2015, Swansea. [2019-11-20]. http: // www. bmva. org/bmvc/2015/papers/paper160/index. html.

[9] Li T Y. Optical fiber communication-the state of the art [J]. IEEE Transactions on Communications, 1978, 26(7): 946-955.

[10] Kaushal H, Jain V K, Kar S. Overview of wireless optical communication systems [M]// Optical Networks. New Delhi: Springer India, 2017: 1-39.

[11] French P, Patterson M. Advances in optical imaging, photon migration, and tissue optics [J]. Optics and Photonics News, 1999, 10(10): 93-94.

[12] Zhang J, Zhang L, Zeng F, et al. Development status of airborne 3D imaging lidar systems [J]. Chinese Optics, 2011, 4(3): 213-232.

[13] Saruwatari M. All-optical signal processing for terabit/second optical transmission[J]. IEEE Journal of Selected Topics in Quantum Electronics, 2000, 6(6): 1363-1374.

[14] Brzozowski L, Sargent E H. Optical signal processing using nonlinear distributed feedback structures [J]. IEEE Journal of Quantum Electronics, 2000, 36(5): 550-555.

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

[16] Ribeiro A, Ruocco A, Vanacker L, et al. Demonstration of a 4×4-port universal linear circuit [J]. Optica, 2016, 3(12): 1348-1357.

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

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

[19] Bagherian H, Skirlo S, Shen Y C, et al. On-chip optical convolutional neural networks[J/OL]. [2019-11-20]. https://arxiv.org/abs/1808.03303.

[20] Li L, Fang N, Wang L T, et al. Research progress in hardware implementations of reservoir computing [J]. Laser & Optoelectronics Progress, 2017, 54(8): 080005.
李磊, 方捻, 王陆唐, 等. 储备池计算硬件实现方案研究进展[J]. 激光与光电子学进展, 2017, 54(8): 080005.

[21] Bao X R. Research progress in optoelectronic reservoir computing system [J]. Laser & Optoelectronics Progress, 2015, 52(3): 030005.
包秀荣. 光电储备池计算系统研究进展[J]. 激光与光电子学进展, 2015, 52(3): 030005.

[22] Zhao Q C, Yin H X. Research progress of reservoir computing using chaotic laser [J]. Laser & Optoelectronics Progress, 2013, 50(3): 030003.
赵清春, 殷洪玺. 混沌光子储备池计算研究进展[J]. 激光与光电子学进展, 2013, 50(3): 030003.

[23] Vandoorne K, Dambre J, Verstraeten D, et al. Parallel reservoir computing using optical amplifiers [J]. IEEE Transactions on Neural Networks, 2011, 22(9): 1469-1481.

[24] Bueno J, Maktoobi S, Froehly L, et al. Reinforcement learning in a large-scale photonic recurrent neural network[J]. Optica, 2018, 5(6): 756-760.

[25] Paquot Y, Duport F, Smerieri A, et al. Optoelectronic reservoir computing [J]. Scientific Reports, 2012, 2: 287.

[26] Duport F, Schneider B, Smerieri A, et al. All-optical reservoir computing[J]. Optics Express, 2012, 20(20): 22783-22795.

[27] Zhang H, Feng X, Li B X, et al. Integrated photonic reservoir computing based on hierarchical time-multiplexing structure[J]. Optics Express, 2014, 22(25): 31356.

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

[29] Shastri B J, Nahmias M A, Tait A N, et al. Spike processing with a graphene excitable laser [J].

Scientific Reports, 2016, 6: 19126.

[30] Peng H T, Nahmias M A, de Lima T F, et al. Neuromorphic photonic integrated circuits[J]. IEEE Journal of Selected Topics in Quantum Electronics, 2018, 24(6): 1-15.

[31] 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.

[32] Bohte S M, Kok J N, LaPoutré H. Error-backpropagation in temporally encoded networks of spiking neurons[J]. Neurocomputing, 2002, 48(1/2/3/4): 17-37.

[33] Gütig R, Sompolinsky H. The tempotron: a neuron that learns spike timing-based decisions[J]. Nature Neuroscience, 2006, 9(3): 420-428.

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

[35] Ponulak F, Kasiński A. Supervised learning in spiking neural networks with ReSuMe: sequence learning, classification, and spike shifting[J]. Neural Computation, 2010, 22(2): 467-510.

[36] Mohemmed A, Schliebs S, Matsuda S, et al. Span: spike pattern association neuron for learning spatio-temporal spike patterns[J]. International Journal of Neural Systems, 2012, 22(4): 1250012.

[37] Yu Q, Tang H J, Tan K C, et al. Precise-spike-driven synaptic plasticity: learning hetero-association of spatiotemporal spike patterns[J]. PLoS One, 2013, 8(11): e78318.

[38] 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.

[39] Zhang T, Wang J, Dan Y H, et al. Efficient training and design of photonic neural network through neuroevolution[J]. Optics Express, 2019, 27(26): 37150.

[40] Xie L, Liang H N, Tong L. Recognition of signal modulation types in complex optical networks[J]. Laser Journal, 2018, 39(7): 130-133.
解琳, 梁海楠, 佟璐. 复杂光网络信号调制类型的识别研究[J]. 激光杂志, 2018,39(7): 130-133.

[41] Pei Y H, Qu Y, Li J M, et al. Modulation recognition of MPSK and MQAM signals based on AlexNet[J]. Laser Journal, 2018, 39(10): 75-78.
裴禹豪, 曲毅, 李锦明, 等. 基于 AlexNet 网络的 MPSK 与 MQAM 类信号的调制识别[J]. 激光杂志, 2018, 39(10): 75-78.

[42] Selden A C. Pulse transmission through a saturable absorber[J]. British Journal of Applied Physics, 1967, 18(6): 743-748.

[43] Dejonckheere A, Duport F, Smerieri A, et al. All-optical reservoir computer based on saturation of absorption[J]. Optics Express, 2014, 22(9): 10868.

[44] Gao Y C, Zhang X R, Li Y L, et al. Saturable absorption and reverse saturable absorption in platinum nanoparticles[J]. Optics Communications, 2005, 251(4/5/6): 429-433.

[45] Williamson I A D, Hughes T W, Minkov M, et al. Reprogrammable electro-optic nonlinear activation functions for optical neural networks[J]. IEEE Journal of Selected Topics in Quantum Electronics, 2020, 26(1): 1-12.

[46] 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.

[47] Bao Q L, Zhang H, Ni Z H, et al. Monolayer graphene as a saturable absorber in a mode-locked laser[J]. Nano Research, 2011, 4(3): 297-307.

[48] Lim G K, Chen Z L, Clark J, et al. Giant broadband nonlinear optical absorptionresponse in dispersed graphene single sheets[J]. Nature Photonics, 2011, 5(9): 554-560.

[49] Hu X, Wang AD, Zeng M Q, et al. Graphene-assisted multiple-input high-base optical computing[J]. Scientific Reports, 2016, 6: 32911.

[50] Yadav R K, Aneesh J, Sharma R, et al. Designing hybrids of graphene oxide and gold nanoparticles for nonlinear optical response[J]. Physical Review Applied, 2018, 9(4): 044043.

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

[52] Ramachandran P, Zoph B, Le Q V. Searching for activation functions[J]. [2019-11-20]. https://arxiv.org/abs/1710.05941.pdf.

[53] 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.

[54] Zang Y B, Chen M H, Yang S G, et al. Electro-optical neural networks based on time-stretch method[J]. IEEE Journal of Selected Topics in Quantum Electronics, 2020, 26(1): 1-10.

[55] Yu Z M, Zhao X, Yang S G, et al. Binarized coherent optical receiver based on opto-electronic neural network[J]. IEEE Journal of Selected Topics in Quantum Electronics, 2020, 26(1): 1-9.