用于人工智能的硅基光电子芯片

Silicon Photonic Chips for Artificial Intelligence

摘要：提出了利用硅基光电子芯片进行人工神经网络计算处理的方法。硅基光电子芯片凭借光子的独特性质，能够在人工神经网络的计算处理中发挥高带宽、低时延等优势。在处理深度学习中大量的矩阵计算的乘加任务时，硅基光电子芯片拥有更高的处理速度和更低的能耗，从而有利于深度学习中的人工神经网络计算速度和性能的提升。

关键词：人工神经网络；硅基光电子芯片；人工智能；深度学习

Abstract: Silicon photonic chips are used to perform artificial neural network computation. Because of the unique properties of photons, silicon photonic chips have the advantages of high bandwidth and low delay in the computation and processing of artificial neural network. When dealing with the multiplication and addition task of a large number of matrix calculations in deep learning, silicon photonic chips have higher processing speed and lower energy consumption, which is beneficial to the improvement of the computational speed and performance of artificial neural network in deep learning.

Keywords: artificial neural network; silicon photonic chips; artificial intelligence; deep learning

白冰 / BAI Bing
裴丽 / PEI Li
左晓燕 /ZUO Xiaoyan

（北京交通大学，中国 北京 100044）
(Beijing Jiaotong University, Beijing 100044, China)

DOI：10.12142/ZTETJ.202101015
网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.TN.20200421.1831.002.html

网络出版日期：2020-04-22
收稿日期：2020-02-20

人工智能发展的着眼点之一是强大的大型数据集处理工具。这就要求计算机在没有获得明确指令的条件下，能快速高效地学习并组合分析大量信息。人工神经网络就是可以进行学习的数据处理计算机，而以人工神经网络为基础的深度学习算法因其在图像识别、问题决策、语言翻译、自动驾驶[1]、医疗辅助[2]等方面的应用而受到学术界和工业界的关注。

目前，人工神经网络几乎全部依赖于传统的电域集成芯片，包括中央处理器（CPU）、图形处理器（GPU）、现场可编程门阵列（FPGA）、专用集成电路（ASIC）。微电子芯片因其结构上无法规避的缺陷，在处理大量的矩阵运算时，面临带宽低、功耗大、速度慢等问题，但人工神经网络实现的基础就是大量的矩阵运算；因此，要想实现深度的人工神经网络，就需要更多的时间和能耗成本。我们可以通过不断提高芯片集成度，进行存内计算等方法解决这个问题；但与此同时，晶体管尺寸不断缩小，晶体管的性能也越来越受到量子效应的影响，这限制了集成度的不断提高。另外，存内计算的方法与现有的人工神经网络算法匹配度不高也限制了存内计算这种方法的应用。

为了解决上述问题，学术界和工业界越来越多地致力于开发新的硬件架构，以适应人工神经网络和深度学习的应用。借助光子器件优势（带宽大、速度快），业界提出将一部分信息承载和计算处理用于改善电域芯片存在的问题。相比于传统的三五族或铌酸锂光器件，硅基光电子芯片上的光器件集成在同一硅衬底上，集成度更好且基本与成熟的互补金属氧化物半导体（COMS）工艺兼容。

利用光电子技术实现的人工神经网络主要包括前馈神经网络（FNN）、循环神经网络（RNN）、脉冲神经网络（SNN）3种类型。马赫·曾德尔干涉仪（MZI）和微环谐振器（MRR）具有干涉、谐振等物理特性，可以实现调制器、滤波器等多种器件功能，被广泛地用于通信、传感等领域。目前相对比较完善的、主流的硅基集成通信芯片是基于MZI和MRR的两种类型，因此人工神经网络芯片也主要基于这两种类型。本文中，我们围绕这两种类型对硅基光电子人工智能芯片的进展进行简要阐述，并对未来的发展态势进行展望。

1 利用光网络进行矩阵运算

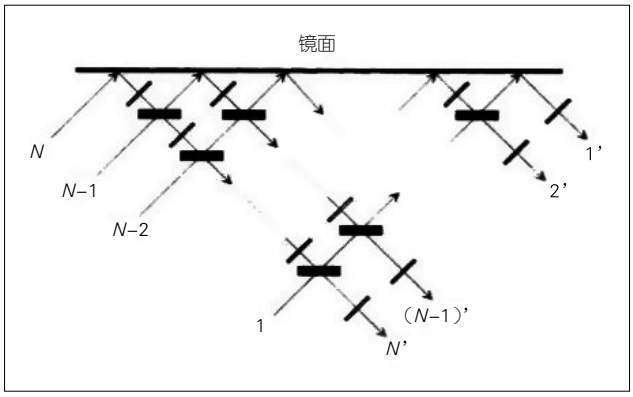
人工神经网络的思路是首先将输入的事物转化为矩阵，然后经过大量的矩阵运算，最终得到所需要的结果。不同算法的处理流程可能会有一些差异，但是都会包含大量的矩阵运算。矩阵运算的基础就是乘积累加运算（MAC）。在光子领域实现MAC操作并不会在本质上消耗能量，这是光子集成电路的优势之一。

1.1 集成 MZI 进行矩阵运算的原理

利用 MZI 进行片上矩阵运算的原理是基于 M. RECK 等于 1994 年提出的酉矩阵分解方法[3]的。在该方法中，可调反射率和透过率的分束器和可调的移相器组成基本单元，并通过电压控制分束器的分光比和移相器的相位实现控制输出端口的光强，如图 1 所示[3]。酉矩阵运算中输入的 $N\times1$ 列向量元素大小由输入光强大小来表示，位置由输入端口的位置表示，输入的多束光分别从入口端 MZI 的一个臂进入 MZI 阵列，$N\times N$ 酉矩阵中的元素使用 MZI 阵列中每一个 MZI 包含的 2 个移相器和分光器的参数来表示。这使得光通过这些 MZI 时，相位和幅度会发生改变，进而达到计算效果。最后根据输出端的 $N\times1$ 个输出光强大小来计算结果列向量元素大小，元素位置由输出端口的位置表示。在进行输入光强的调制和输出光强的探测后，利用光网络可实现酉矩阵的计算。

图 1 为光路酉矩阵分解结构。其中，较大黑横矩形表示分光比可调分束器，小黑斜矩形表示移相器，上方细长黑矩形为全反射镜面。

在酉矩阵实现之后，我们可以利用奇异值分解（SVD）的方法对任意矩阵进行分解，即 SVD 将矩阵分解为 2 个酉矩阵和 1 个对角矩阵酉矩阵，光路对角矩阵的模拟用衰减器即可完成，MZI 也可以做衰减器。这样就实现了利用 MZI 进行矩阵运算。

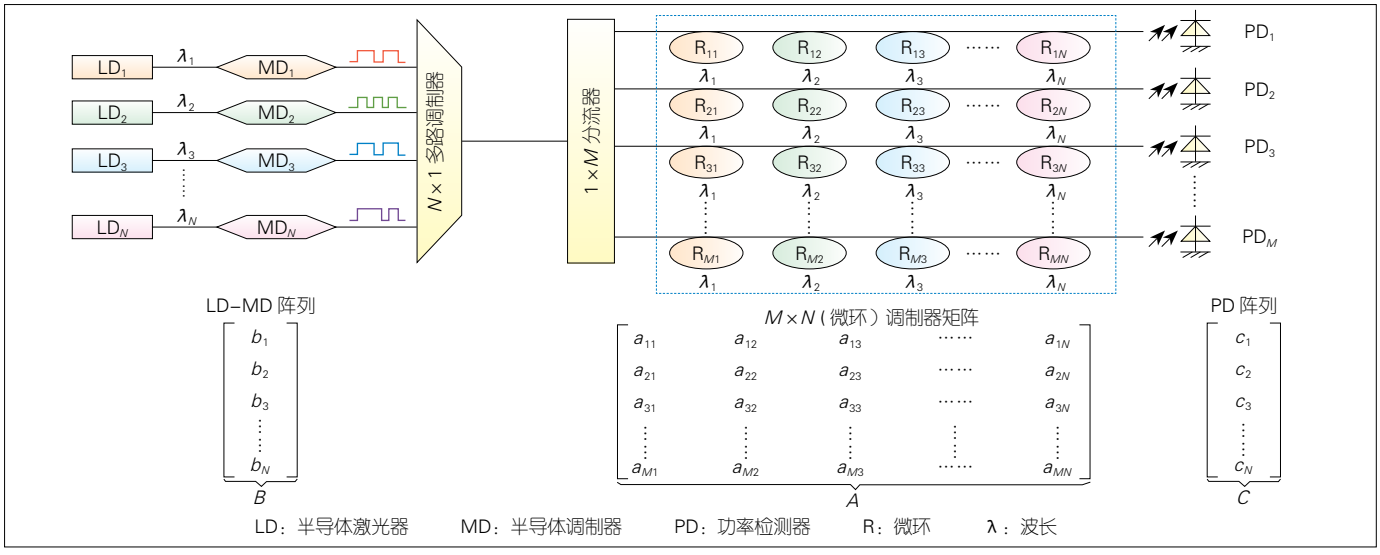


▲图 1 光路酉矩阵分解结构

1.2 集成 MRR 实现矩阵运算的原理

MRR 可以先将特定波长的光信号耦合到环上进行调制，然后再耦合进直波导。MRR 实现矩阵运算的原理为：通过透过率的调节来实现矩阵的表示。首先矩阵运算中输入的 $N\times1$ 列向量中的元素大小用光强大小表示，列向量中元素的位置由不同的波长表示（因为输入列向量来自于电域信号，所以需要通过调制器进行电光转换）。$M\times N$ 矩阵的每一列元素用同一个波长表示，不同列用不同的波长表示，也就是说同一列的 MRR 耦合的是同一个波长。矩阵的每一行用一个公共波导以及耦合在其上的 MRR 表示，然后每行上的 MRR 根据谐振波长的不同，分别对输入的不同波长的光信号进行强度调制以实现乘法器，强度的大小表示的是此行元素的大小，然后 MRR 再将不同波长的光信号耦合入公共的波导实现加法器。最后矩阵运算结果是一个 $M\times1$ 列向量，其元素大小通过光强大小来表示，然后经光电探测器进行光电转换后，再通过测量电流大小后得到。

图 2 所示的是 YANG L. 等提出的一种利用 MRR 来实现矩阵运算的方法[4]。这种 MRR 光网络结构可以执行一个 $M\times N$ 矩阵 $\boldsymbol{A}$ 和一个 $N\times1$ 向量 $\boldsymbol{B}$ 的乘法。$\boldsymbol{B}$ 是输入向量，用 N 个不同波长光信号的光功率大小来表示向量 $\boldsymbol{B}$ 中的元素。这一个列向量 $\boldsymbol{B}$ 是通过 N 个外部调制或直接调制激光二极管



▲图 2 微环谐振器实现矩阵运算的结构

所生成的。N 个光信号通过一个多路复用器被多路复用到一个公共波导上，然后通过一个 $1\times M$ 的光分路器将其平行投影到 M 行调制器上。矩阵 $\boldsymbol{A}$ 的 a_{ij} 元素由位于矩阵第 i 行和第 j 列的 MRR 的透射率表示，位于同一行的每个 MRR 仅对具有特定波长的光信号进行操作。随着 $M\times N$ 光脉冲通过 MRR 矩阵，光信号就在环上进行了所有的 $M\times N$ 乘法过程。在 M 个环上进行乘法运算后，其累加过程在公共输出波导中进行。因为不同波长的信号在公共波导中几乎不会互相干扰。结果向量 $\boldsymbol{C}$ 的元素由光检测器阵列检测到的 M 个光功率表示。由此，利用 MRR 进行的矩阵运算便得以实现。

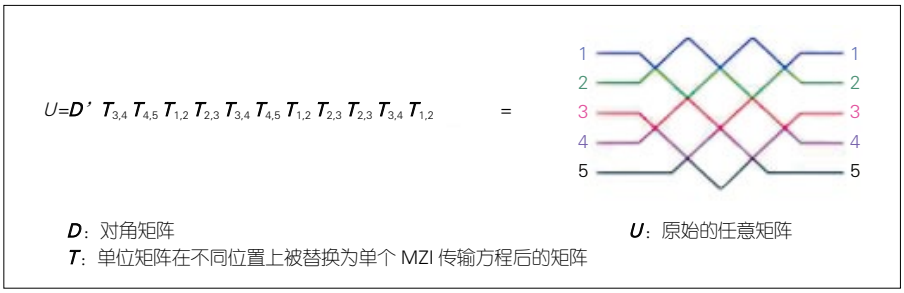
2 现阶段片上人工神经网络的实现

由光来进行矩阵运算是解决人工神经网络需要大量矩阵运算的思路之一。现阶段，硅基集成的、可用于搭建人工神经网络的典型基础器件就是 MZI 和 MRR。通过这些器件的光学特性实现了 MAC 运算和脉冲神经元的模拟。借助光子数据处理方面的优势，我们将软件和硬件进行深度匹配，使用高效的光电计算取代微电子处理器的计算 [5]。光电子集成、数学和软件算法等领域的深度交叉是解决人工神经网络算法大量密集计算问题的路径之一，也是人工神经网络算法片上实现的发展趋向。

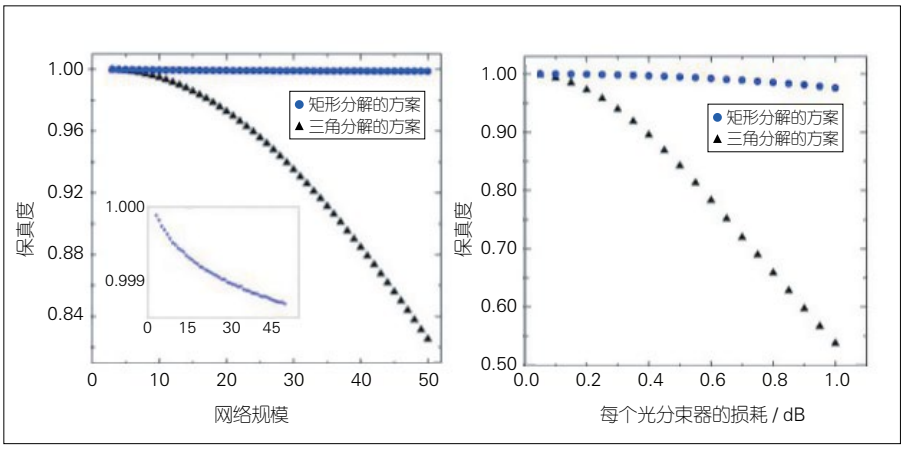
2.1 MZI 型片上人工神经网络

在基于 MZI 构建的前馈人工神经网络中，信息从输入层单向传递到输出层。信号前向传播时，不需要将输出再次反馈，只需要进行加、乘，以及比较操作即可，这与擅长矩阵运算的光网络相匹配；因此，此种方法的光路硬件的实现获得了广泛的探索和关注。虽然 M. RECK 等发现以 MZI 进行酉矩阵的分解方法时并未考虑集成 [3]，但是 MZI 型人工神经网络日渐向集成发展。W. R. CLEMENTS 等在 2016 年基于 M. RECK 等的三角分解法提出了矩形分解法 [6]，将 MZI 进行重新排布来实现酉矩阵运算。通过将 MZI 的排布形状从三角转化为矩形，减少一半的光学深度，同时也增加了计算网络的误差容忍度。酉矩阵的分解过程如图 3 所示。酉矩阵矩形分解方法比酉矩阵三角分解方法更有优势。这是因为酉矩阵三角分解方法的光路是不对称的，从而导致了一些传输过程中的误差。矩形设计减小了线路不对称性，并缩短了最长链路的长度，从而减少了光传播的路径损耗和误差。在对 500 个随机生成含误差酉矩阵传输的模拟中，随着酉矩阵规模 N 从 2 扩大到 50，三角分解方法的准确度由 100% 下降到约 82%；但是矩形分解方法的准确度并未发生明显下降，一直保持在约 100%，具体如图 4 所示。

2017 年，SHEN Y. C. 等利用 56 个 MZI 实现了可以用于元音识别的全连接片上神经网络，制成了光子干涉单元芯片。芯片的部分结构如图 5 所示 [7]。在这个设计中，通过 MZI 阵列进行神经元线性部分的运算，人工神经网络中的非线性激活函数采用电域仿真的方法得以实现，最终可实现全连接神经网络的片上系统。该芯片搭建了 2 层、每层 4 个神经元的全连接神经网络。图 5 所示的芯片结构只有 1 个酉矩阵和 1 个对角阵，所以应用时要先将元音信号转为光信号，取得结果放到电域中处理为光信号再传进来，至此完成一层计算。将以上过程循环两遍即为 2 层神经网络。上述结构在对 4 个元音的分类的实验中，能够从大量不同元音的语音信号中正确识别和分类元音，准确率达到 76.7%。



▲图 3 酉矩阵的四边形分解过程



▲图 4 三角分解和矩形分解对误差的容忍度

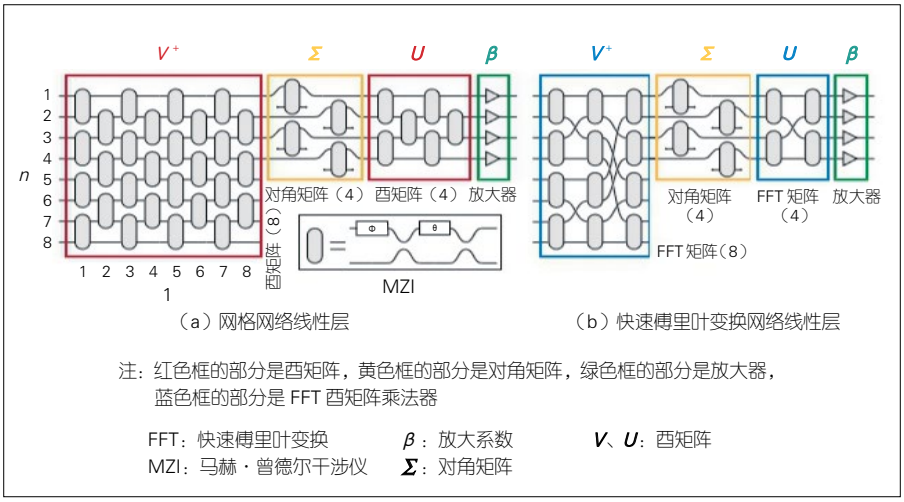
2019 年，M. Y. S. FANG 等研究了两种类型的 MZI 神经网络，分别为 GridNet（网格网络）、FFTNet（快速傅里叶变换网络），其物理结构如图 5 所示[8]。其中，矩阵的分解采用的是 SVD。FFT 酉矩阵乘法器是非通用性的乘法器，它由 Cooley–Tukey FFT 算法启发而来，用牺牲通用性的方式来换取结构上的紧凑性。由图 6 可以看出，GridNet 和四边形结构是同一种结构，它和 FFTNet 结构皆为 8×4 的线性矩阵运算器。这两种结构均为仿真，在零误差的情况下，GridNet 的准确率约为 98%，FFTNet 的准确率约为 95%，因此零误差时的 GridNet 准确率比 FFTNet 的准确率高；但是在有差错的情况下，FFTNet 的容错率要比 GridNet 高。在综合误差从 0 升高到 0.02 时，FTTNet 的准确率由约 95% 降到约 93%，而 GridNet 的准确率由约 97% 降到约 48%。越小的网络差错传播所带来的误差就会越小，这导致了 FFTNet 的稳定性要优于 GridNet。

综上可得，无论是在矩阵规模扩大还是误差增加的情况下，三角结构的准确率低于矩形结构。矩形结构和 GridNet 是同种结构，它和 FFTNet 结构各有利弊：GridNet 结构的通用性好，但在存在误差的情况下，准确率低；FFTNet 通用性差，但在误差存在的情况下，准确率高。

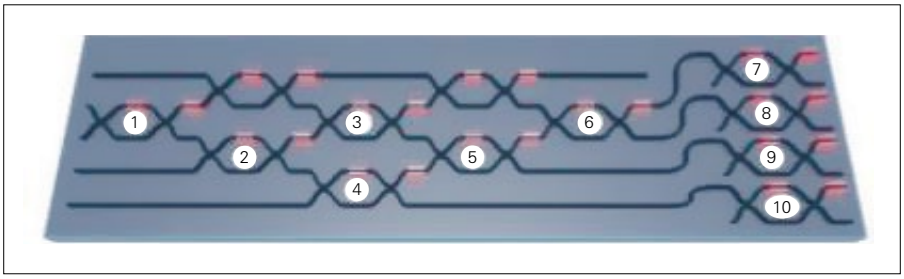
2.2 MRR 型片上人工神经网络

MRR 神经网络主要用来实现脉冲神经网络（SNN）。这种网络考虑了时间信息，相比于 FNN 和 RNN 更加接近于真实的人脑运作情况，被称为第 3 代神经网络[9]。

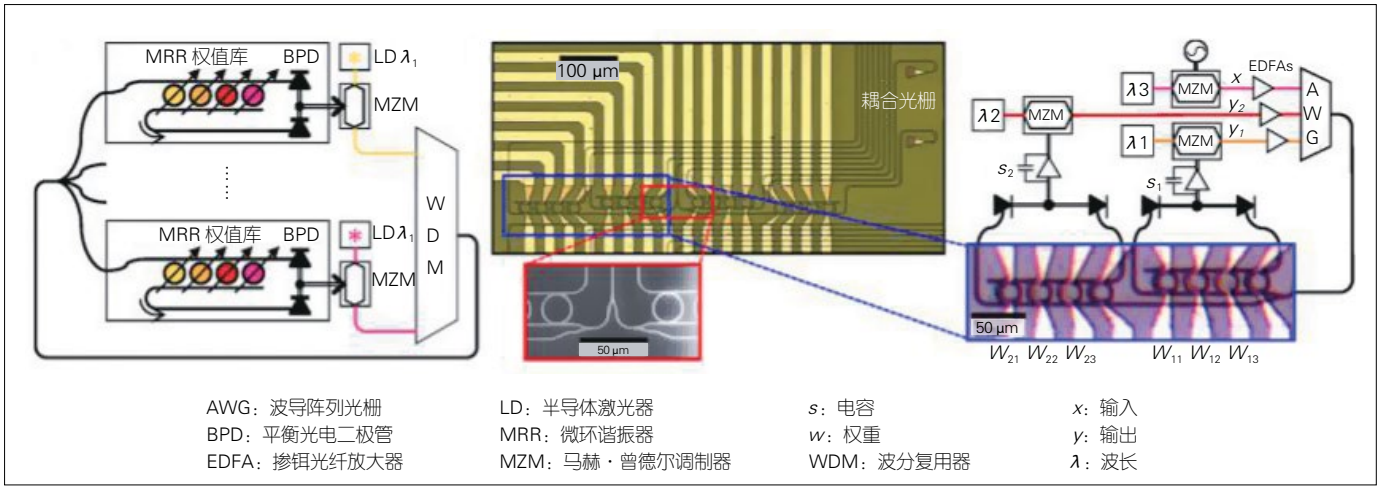
图 7 所示为 A. N. TAIT 等于 2017 年提出的一种广播式 MRR 权值库结构的神经网络[10]。这是一种以 MRR 调制器作为神经元，由 MRR 权值库连接而成的网络。每一个 MMR 都承担着一个权值，每一横条聚集在一起的 MMR 叫权值库。该芯片结构包含 4 个节点，



▲图 5 两种不同类型的模拟片上人工神经网络架构



▲图 6 全连接神经网络芯片结构



▲ 图 7 MRR 权值库结构

带有 16 个 MRR。该结构证明了硅光子电路与连续神经网络模型之间存在数学同构关系。根据这种同构性，我们利用“神经编译器”对一个模拟的 24 节点硅光子神经网络进行编程，完成了微分系统仿真任务。根据推算，与传统的解决相同问题的 CPU 相比，此结构的处理速度将提高 294 倍。

2019 年，A. N. TAIT 等提出了一种神经拟态的片上结构[11]，该结构主要由 2 个光探测器和 1 个 MRR 组成，如图 8（a）所示。神经元阵列由电脉冲强度调控单元及延时单元构成，除泵浦激光器外，整体网络可实现片上集成。每个 MRR 只有一个波长（λ_i）。该结构将 MRR 强度调制器和平衡的光电探测器组成电光脉冲强度和延迟调控单元，并使用电光脉冲进行调控，以实现复杂的脉冲神经网络。当输入为 2 ns 脉冲偶极子，第 2 次的输入相比于第 1 次输入延迟一个波长，进而产生 t = 0 时的脉冲重合。据此测得的加强、饱和、抑制 3 种情况下的结果如图 8（b）所示。在图 8（b）中，在可见增强情况下，脉冲出现过冲现象（超过 57%）；而在饱和情况下，脉冲只为输入脉冲和的 56%，且单脉冲抑制不完全。虽然存在一些问题，但是这种结构形成了全光广播权值神经网络的组件类型，在一个集成的光子组件中包含了光到光的非线性、扇入和非确定性级联，实现了光子神经元的网络兼容的能力。

目前 MRR 人工神经网络偏向于贴合神经网络的数学同构模型的研究，采用比较统一的采用扇入结构、光电光转换等实现方案。

2.3 MZI 型与 MRR 型片上人工神经网络的对比

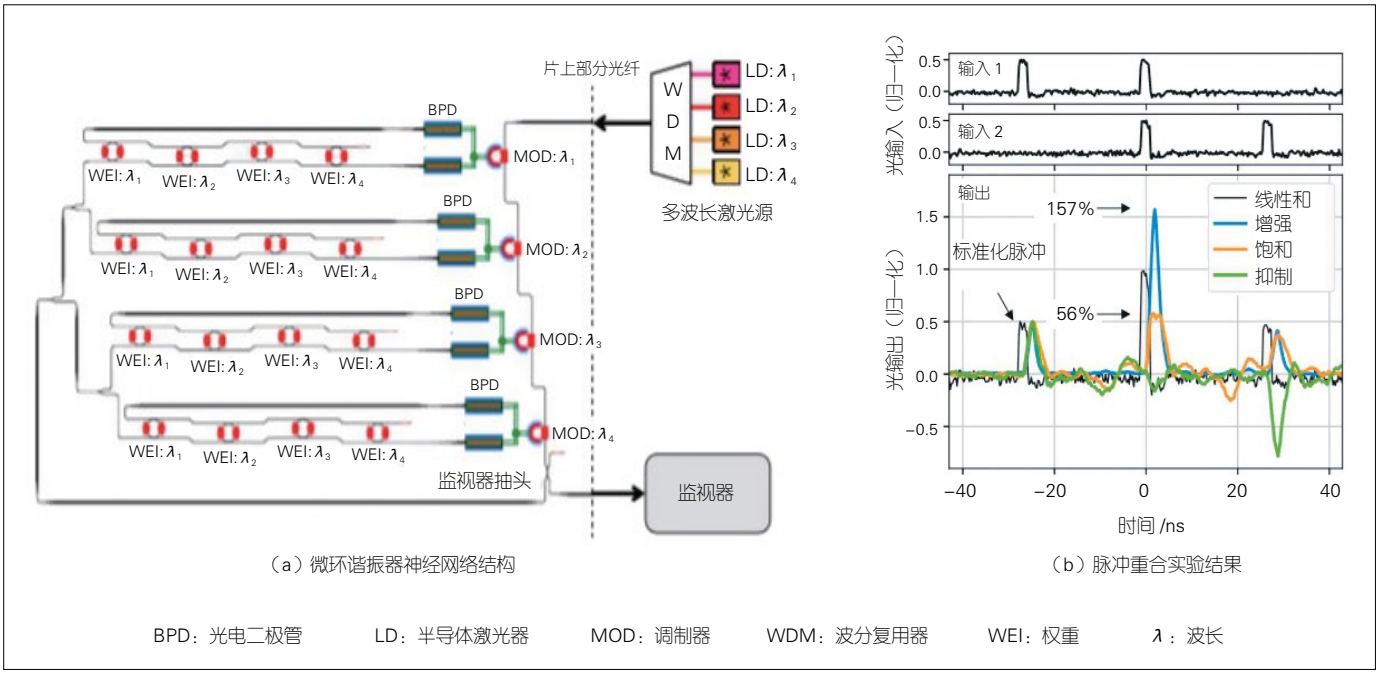
MZI 型片上人工神经网络是根据酉矩阵分解和计算来设计数学同构性的；而 MRR 型片上人工神经网络则直接以普通矩阵的计算来设计数学同构性。两者本质上都是用光器件来表现数学计算。由于 MRR 型仍在结构探索阶段，相比于 MZI 型，准确性远远不足；而 MZI 型有较为成熟的应用测试，但相比于传统芯片，准确性仍有不足。

3 片上人工神经网络实现方案面临的挑战

利用光子进行计算具有诸多优势，但目前仍存在一些问题：

（1）非线性激活函数是用来增加神经网络非线性的一种 S 形状的函数，它的硬件实现起来比较困难。现在非线性函数的实现方式分为两种：一种就是转换到电域再处理[7]或利用电域辅助处理[12]；另一种就是利用特殊材料，如可饱和吸收体和石墨烯等进行处理。一方面，光电转换限制了数据处理速度的进一步提升；另一方面，大部分特殊材料的片上集成较为困难，不能与互补金属氧化物半导体（CMOS）工艺兼容。

（2）MZI 的长度约为 200 μm，MRR 的长度约为 25 μm。相比于电域的器件，芯片集成度差，目前工艺方面还有进一步提升的空间。虽然看起来 MRR 要比 MZI 小一些，但是它们基本都属于一个数量级。另外，由于



（a）微环谐振器神经网络结构　　（b）脉冲重合实验结果

BPD：光电二极管　LD：半导体激光器　MOD：调制器　WDM：波分复用器　WEI：权重　λ：波长

▲ 图 8 微环谐振器神经网络及其实验效果

目前硅基集成光器件的工艺仍旧不够成熟，器件的一致性、稳定性较差。

（3）目前我们需要对光电人工智能芯片的匹配算法和外围电路进行设计[13]，并需要将各领域技术深度融合。这个结合的过程需要重新进行布局和设计，在目前没有统一标准的情况下。每一个光网络结构的出现都可能会导致外围匹配的电路和算法重新被调整和优化。

4 结束语

光电神经网络能够利用光子技术的优点并配合外围电域进行处理，在提升计算速度的同时也可以降低运行功耗。无论是基于 FNN 的 MZI 前向人工神经网络芯片，还是基于 SNN 的 MRR 神经拟态人工神经网络芯片，都可以利用硅基光电子技术进行实现。此外，光电神经网络亦会随着硅基光电子技术的成熟而不断取得突破，如硅基片上光源、放大器、硅基单片集成、硅基新材料融合等新型硅基光电子技术，都将为光电神经网络的物理研究提供崭新的、开阔的思路。同时，随着与光电神经网络相匹配的算法演进，相信在将来的研究中，硅基光电子技术、硅基光电子芯片将为人工智能领域带来全新的技术架构和重大的产业升级。

参考文献

[1] 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). Los Angeles, USA: IEEE, 2017. DOI:10.1109/ivs.2017.7995703

[2] ESTEVA A, KUPREL B, NOVOA R A, et al. Dermatologist-level classification of skin cancer with deep neural networks [J]. Nature, 2017, 542(7639): 115-118. DOI: 10.1038/nature21056

[3] RECK M, ZEILINGER A, BERNSTEIN H J, et al. Experimental realization of any discrete unitary operator [J]. Physical review letters, 1994, 73(1): 58. DOI:10.1103/physrevlett.73.58 DOI:10.1103/physrevlett.73.58

[4] YANG L, JI R Q, ZHANG L, et al. On-chip CMOS-compatible optical signal processor [J]. Optics express, 2012, 20(12): 13560. DOI:10.1364/oe.20.013560

[5] 白冰，赵斌，杨钊 . 一种光子神经网络芯片以及数据处理系统 : CN110503196A [P]. 2019-11-26

[6] CLEMENTS W R, HUMPHREYS P C, METCALF B J, et al. Optimal design for universal multiport interferometers [J]. Optica, 2016, 3(12): 1460. DOI:10.1364/optica.3.001460

[7] SHEN Y C, HARRIS N C, SKIRLO S, et al. Deep learning with coherent nanophotonic circuits [C]//2017 IEEE Photonics Society Summer Topical Meeting Series (SUM). San Juan, USA: IEEE, 2017: 441-447. DOI:10.1109/phosst.2017.8012714

[8] FANG M Y S, MANIPATRUNI S, WIERZYNSKI C, et al. Design of optical neural networks with component imprecisions [J]. Optics express, 2019, 27(10): 14009. DOI:10.1364/oe.27.014009

[9] MAASS W. Networks of spiking neurons: the third generation of neural network models [J]. Neural networks, 1997, 10(9): 1659-1671. DOI: 10.1016/s0893-6080(97)00011-7

[10] 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. DOI: 10.1038/s41598-017-07754-z

[11] TAIT A N, DE FERREIRA L T, NAHMIAS M A, et al. Silicon photonic modulator neuron [J]. Physical review applied, 2019, 11(6): 064043. DOI: 10.1103/physrevapplied.11.064043

[12] 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. DOI:10.1109/jstqe.2019.2930455

[13] 白冰，赵斌，杨钊 . 一种计算电路以及数据运算方法 : CN110597756A [P]. 2019

作 者 简 介

白冰，北京交通大学光波技术研究所在读博士研究生；主要从事硅光集成计算器件、光电异构计算架构和光电融合神经网络算法等领域的研究；已申请专利 12 项。

裴丽，北京交通大学教授、博士生导师；主要从事全光交换、特种光纤、光电器件及基于智能光纤传感的物联网的研究；主持科研项目 10 余项，发表 SCI、EI 论文 200 余篇。

左晓燕，北京交通大学光波技术研究所在读博士研究生；主要从事光神经网络、光电器件领域的研究。