高算力光学张量卷积运算芯片基础研究

张文甲
（上海交通大学光子传输与通信全国重点实验室，上海 200240）

摘 要： 卷积神经网络是计算机视觉和目标检测等领域应用最成功的算法之一. 随着高清图像和视频等数据爆发式增长，智能处理芯片需要更强的算力和更小的功耗. 光子技术的多维特征和波动物理模型为高算力张量卷积运算提供了物理基础，有望从根本上突破电芯片在提升算力和降低功耗上不可逾越的物理限制. 本文介绍高算力光学张量卷积运算芯片基础研究的研究动机、主要研究挑战与解决思路及未来展望，探讨限制光学张量卷积运算应用的主要因素，推动光学张量卷积计算从基础研究走向大规模应用.

关键词： 光学卷积神经网络；光学张量卷积；张量计算；光学神经网络
基金项目： 国家自然科学基金（No.62235011）
中图分类号： TP381；TP183 **文献标识码：** A **文章编号：** 0372-2112(2025)04-1361-04
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.12263/DZXB.20240294

Fundamental Research of High Computation-Capability Devices for Optical Tensor Convolution Operation

ZHANG Wen-jia
(*State Key Laboratory of Photonics and Communications*, *Shanghai Jiao Tong University*, *Shanghai* 200240, *China*)

Abstract: Convolutional Neural Network is one of the most successful algorithms in the fields of computer vision and object detection. With the explosive bandwidth growth of high-definition images and videos, intelligent computing processors require higher computational capability with less power consumption. Photonic technology has inherent capability of coherent combination and multidimensional manipulation, and will become an inevitable approach to realize tensor convolution operations. This paper introduces the research motivation, primary research challenges, solution approaches, and future prospects of high computation-capability optical tensor convolutional devices. It also explores the main limiting factors restraining the application of optical tensor convolution operations, aiming to drive this promising technology from basic research to large-scale applications.

Key words: optical convolutional neural network; optical tensor convolution; tensor operation; optical neural network
Foundation Item(s): National Natural Science Foundation of China (No.62235011)

1 研究动机

2024年国务院政府工作报告指出：“深化大数据、人工智能等研发应用，开展‘人工智能+’行动，打造具有国际竞争力的数字产业集群”. 其中，“人工智能+”行动准确体现了人工智能技术赋能百业的客观现实，将对传统的技术形态、商业模式和生活方式产生重大且深远的影响. 近十年人工智能（Artificial Intelligence，AI）技术的高速发展对微电子芯片处理能力提出了更高需求. 据OpenAI统计，AI算法的算力需求增长速度为3.43个月翻一番；用于芯片间互连的Serdes串行速率每3～4年翻一番，2024年已经达到200 Gbps；图形处理芯片（Graphics Processing Unit，GPU）的算力十年增长近200倍，GPU性能预计每2年翻一番. 然而，日益增长的人工智能应用与迭代趋缓的CMOS芯片技术形成了鲜明的发展矛盾. 此外，中国在先进芯片制程工艺方面严重受限，更加需要广大研究人员发挥主观能动性，而光计算是实现高算力高能效智能计算处理器的重要手段.

为什么光计算是一条可行技术路线？主要原因有两个[1]. 第一，光的多维特征为高速张量化模拟运算提

收稿日期：2024-04-02；修回日期：2025-05-08；责任编辑：宋甲英

供了物理基础. 光子工作频率在100 THz以上,具备几十THz级的工作带宽、fs级脉冲宽度和超低损耗的传输能力,具有偏振、幅度、波长和相位等多维特征. 此外,光子技术是目前高速通信最主要的技术手段,具备从基础器件、封装到模块的完整产业链. 第二,光学波动物理模型蕴含着矩阵运算和卷积运算等基础计算操作的数理基础. 光学结构,如光学透镜,约束了光子在特定时空条件下的行为模型. 而半导体制造技术的快速进步推动了介观尺度光子学的革命,为大规模光计算提供了更小体积、更高效率的光电器件,使探索先进光智能计算处理器成为可能[2]. 综上,利用多维光学物理特性以及设计光学微结构来操控和约束光学行为,构建专用光物理加速运算单元,才能真正发挥光计算独特的优越性.

基于以上两点考虑,我们开展高算力光学张量卷积运算芯片基础研究[3~5],其基本逻辑是将张量卷积运算映射到多维光芯片中. 卷积神经网络的设计中,其生物学机理是来自诺贝尔生理学或医学奖的大卫·休伯尔(David H. Hubel)和托斯坦·维厄瑟尔(Torsten N. Wiesel)关于视觉信息处理的研究成果,而基本操作是生物视觉系统中的二维卷积运算,其物理本质就是光学系统二维卷积运算. 在计算机视觉中,输入彩色图像和卷积核均是张量数据. 由于当前计算机受到冯·诺依曼构架限制,张量卷积运算过程被拆分成大量的矩阵乘加运算和缓存读写. 这是造成张量卷积运算功耗、时延和算力瓶颈的根本原因. 光学系统实现二维卷积的过程可以完全不同,光学透镜从物理上就实现了二维卷积运算. 为了能够适配目前存储系统串行结构并发挥集成光子学优势,我们利用色散效应将光学时空二维卷积运算映射到时频域中,将空间的二维透镜映射成时间透镜. 此外,利用波段复用、空间复用等多种复用方法,二维卷积运算可以进一步扩展为高维张量卷积运算. 这样,光学张量卷积运算将复杂的数学运算映射到光学物理维度和空间结构之中,在光速飞行中完成运算.

因此,光学张量卷积运算芯片基础研究主要面向光学张量卷积运算新器件,建立张量卷积运算数学原理与多维光学变换之间的映射关系,解决张量卷积运算光学降维映射方法、突破张量卷积运算速度极限的器件机制和模数协同光学卷积神经网络构架3个关键科学问题,目标是建立从芯片设计到目标识别应用准确率评估的光学卷积神经网络仿真平台,研制高算力可重构张量卷积运算芯片,最终实现面向高通量图像目标识别应用的芯片性能评估.

2 研究挑战与解决思路

虽然近年来光学卷积运算芯片方案层出不穷,但是绝大多数演示系统采用光学矩阵乘加运算来模仿电域卷积运算实现方式. 光学张量卷积运算芯片和系统研究处于初级阶段,更加深入的光学张量卷积运算数学原理、芯片物理基础和光电模数协同构架等基础性问题还有待解决[6]. 从Web of Science的智能光电计算研究论文关键词统计来看,从2020年起,集成化和精度两大主题均超过50%,足以得出光学计算的主要研究挑战和困难点. 我们认为,光学张量卷积运算的研究必须面向高维图形目标识别应用,以协处理器形式融入计算构架之中,以此为目标,主要面临3个挑战.

第一,如何实现高算力? 为了提高算力,光计算芯片需要增大规模和调制速度. 然而,由于硅光芯片无法实现片内增益,因此规模越大,芯片插损也越大,最终影响输出计算精度. 调制速度的提高意味着需要更加大采样率的模拟数字转换电路,进而增加系统设计难度和实现功耗. 高算力、芯片规模、计算精度和计算功耗等4个计算芯片关键指标,在模拟计算和放大受限的前提下需要精准设计与协调,更需要构建系统级各物理参数管理来进行计算性能优化. 我们主要利用光学器件在时频空的多维调控和复用能力,借助高速调制和波分复用等光通信领域成功的技术积累,构建光域张量卷积运算方法和张量卷积核芯片,通过色散介质设计提供卷积运算所需的光延迟,从而极大地降低数模转换和缓存读写的数量.

第二,如何实现深度和可重构的卷积神经网络? 光学张量卷积运算芯片可以完成一级张量卷积运算,但是这只是卷积神经网络的一部分. 由于目前未出现实用的全光非线性激活和池化操作. 光域非线性激活和池化即便可行,也还需要考虑深度级联中能量的损耗问题,这又会涉及前述输出信噪比和精度损失. 因此,电域处理器的参与将无法避免,当然电域处理可以是数字形式,也可以是模拟形式[7]. 若采用数字形式,成熟的电域非线性激活和池化算法可以很方便地移植进来,同时也可完成对光信号的电域整形,利于深度和可重构卷积神经网络的设计. 因此,我们认为光电在智能处理器设计中是无法割裂开来的,光计算配套的专有电芯片带来系统设计的便利性,但是也将增加光电转化的能量成本. 光电计算的另一种形态是光和电芯片构成环形结构,单层卷积操作结束之后,通过电域信号重新送到同一光芯片之中. 光电环形结构构建类似环形谐振腔,可以实现任意深度的卷积神经网络.

第三,如何实现高鲁棒的卷积神经网络应用? 高算力和深度都无法绕开模拟计算的精度问题. 冯·诺依曼在《计算机与人脑》一书中指出:“模拟计算机的主要局限性与精度有关”. 而目前数字计算机需要极高的精

度要求以能解决复杂问题,其原因与当前数值程序的固有结构有莫大关系. 在数字计算机中,所有函数都会映射到基本的运算结构,大多数函数的运行都需要相当长的、可能是迭代定义的基本操作序列,因此数字精度可能被冗长的基本计算流程逐渐消耗. 模拟计算则是整体计算而非逐层计算,应该避免多种计算操作的模拟域级联造成数字精度的逐层耗尽. 因此,需要面向目标识别应用的模数协同系统设计,以充分发挥张量卷积运算的算力优势. 我们的主要思路包括两条:首先,采用模数协同的光学张量卷积运算设计方法,将光器件、光电转换器件的非线性模型纳入系统设计中,面向具体标准应用设计光计算定制算法,在精度受限的情况下实现高计算准确度的应用演示;其次,利用现有AI技术来建立误差学习模型分析热噪声和量化噪声的统计特性,掌握光学模拟系统中的光电噪声规律,进而提高模拟计算中的数字精度.

除了以上3个技术挑战,光计算领域还需要关注评价标准的问题. 首先,模拟计算的精度定义应该包括模数转换电路和光模拟计算系统,单个光学器件或者级联器件的高精度设计应该放到数字系统中才能真正体现其计算的价值. 其次,算力能效的评价也需要进一步统一评价标准,光计算能耗组成和算力供给应面向实际应用,放到与AI算力集群相平齐的标准进行对比. 无论是在项目指标还是产品参数等方面,光计算的每秒操作数、矩阵乘加数或者浮点计算能力应该进一步统一,以建立符合光计算实际技术特征的评价参数. 最后,在应用演示方面,应建立统一评价流程和接口,各种高算力光计算方案需在相同的评价体系中进行硬件计算能力的客观评价.

基于上述研究思考,研究团队开展光学张量卷积运算芯片基础研究,坚持“成熟工艺”和“极简构架”两个基本点,以光学二维卷积操作为基础,简并可重构小卷积核和众核芯片为构架,利用集成光互连芯片提高算力,开发容错训练方法来提高光计算可用性. 如图1所示,本研究从卷积运算物理模型和卷积运算光学芯片设计开始,提出可重构张量卷积核和数模协同方法,最终实现光学卷积协处理器和感前后一体化处理系统.

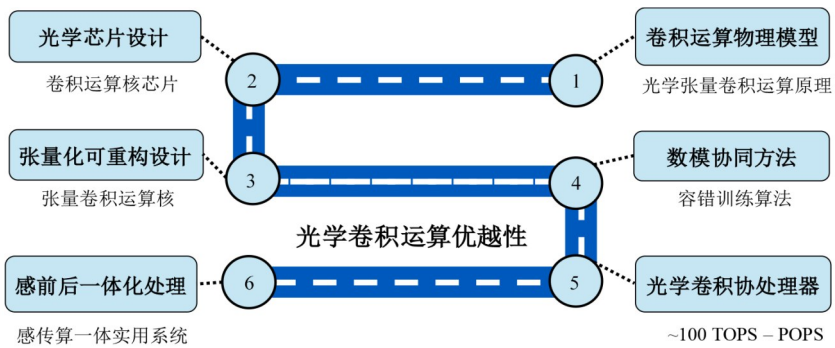

图1　光学张量卷积运算芯片研究路径

3　未来展望

近年来,以光学张量卷积运算为代表的光计算领域受到学术界和投资界热捧,热点论文和初创公司时常活跃于各大宣传平台. 但是,光计算应用场景还比较少,行业面临芯片设计、系统构架、应用接口等方面的痛点. 面向未来,我们认为需坚持“传输即计算、结构即功能”的光计算理念,以“光电一体、数模融合”的设计思路,证明光计算的优越性.

第一,建立光学张量卷积运算设计平台. 光学模拟计算的优势应该体现在系统性能和专有应用上. 因此,需要建立垂直集成的设计仿真平台,从计算芯片、卷积方法、光电转换、训练算法到典型应用形成整体的设计仿真框架和应用预期. 特别是将器件非线性物理特性和噪声类型纳入设计平台之中,评估其对专有应用正面或者负面的影响.

第二,设计光学卷积核微型光电芯粒. 尽可能减少光学卷积运算芯片操作规模,降低芯片操作过程中的噪声积累和能量损耗. 建立以小卷积核为基础的光学卷积运算芯片,设计芯片控制稳定驱动和逻辑电路,实现单一运算功能卷积核微型光电芯粒,以此作为最基本的计算组件.

第三,实现深度可重构的张量卷积神经网络. 基于微型光电芯粒的空间复用集成和组合,配合光电信号处理模块和光电环形控制结构,构建光学张量卷积深度神经网络. 区别于GPU网络结构,光电环形计算结构可以在模块尺度下完成大通量的数据卷积操作,避免大量数据的空间搬移.

第四,构建通用化生物目标识别应用框架. 好用、省电是光学张量卷积运算平台的关键优势. 基于上述光电智能计算系统的优势,拓展光算电协应用类型,使

光协处理器成为通用型智能加速器,创建光电混合智能计算产业化新赛道.

参考文献

[1] MCMAHON P L. The physics of optical computing[J]. Nature Reviews Physics, 2023, 5: 717-734.

[2] SHEKHAR S, BOGAERTS W, CHROSTOWSKI L, et al. Roadmapping the next generation of silicon photonics[J]. Nature Communications, 2024, 15(1): 751.

[3] HUANG Y Y, ZHANG W J, YANG F, et al. Programmable matrix operation with reconfigurable time-wavelength plane manipulation and dispersed time delay[J]. Optics Express, 2019, 27(15): 20456-20467.

[4] JIANG Y, ZHANG W J, YANG F, et al. Photonic convolution neural network based on interleaved time-wavelength modulation[J]. Journal of Lightwave Technology, 2021, 39(14): 4592-4600.

[5] JIANG Y, ZHANG W J, LIU X Y, et al. Physical layer-aware digital-analog co-design for photonic convolution neural network[J]. IEEE Journal of Selected Topics in Quantum Electronics, 2023, 29(6): 7400509.

[6] 张文甲, 姜越, 何祖源. 光子卷积神经网络的研究思考[J]. 中国计算机学会通讯, 2022, 18(7): 62-68.

[7] CHEN Y T, NAZHAMAITI M, XU H, et al. All-analog photoelectronic chip for high-speed vision tasks[J]. Nature, 2023, 623(7985): 48-57.

作者简介

张文甲 男,1984年出生于浙江省余姚市. 现为上海交通大学光子传输与通信全国重点实验室教授,博士生导师. 主要研究方向为集成光计算技术.

E-mail: wenjia.zhang@sjtu.edu.cn