Semiconductor Optical Amplifier-based Photonic Integrated Deep Neural Networks

Citation for published version (APA):
Shi, B. (2022). *Semiconductor Optical Amplifier-based Photonic Integrated Deep Neural Networks*. [Phd Thesis 1 (Research TU/e / Graduation TU/e), Electrical Engineering]. Eindhoven University of Technology.

Document status and date:
Published: 30/06/2022

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

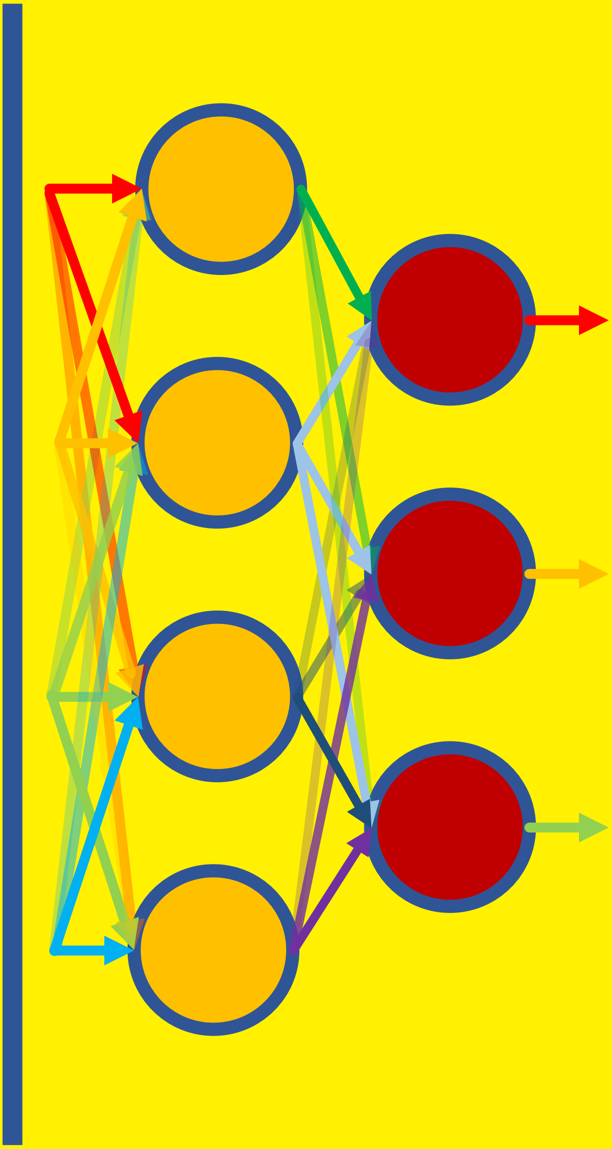
Take down policy
If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

Download date: 04. Oct. 2023

Semiconductor Optical Amplifier-based Photonic Integrated Deep Neural Networks



Bin Shi

Semiconductor Optical Amplifier-based Photonic Integrated Deep Neural Networks

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van de
rector magnificus prof.dr.ir. F.P.T. Baaijens, voor een
commissie aangewezen door het College voor
Promoties, in het openbaar te verdedigen
op donderdag 30 juni 2022 om 13:30 uur

door

Bin Shi

geboren te Wuzhou, China

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

Voorzitter: prof.dr.ir. P.G.M. Baltus
Promotor: dr. R. Stabile
Copromotoren: dr. N. Calabretta
prof.ir. A.M.J. Koonen
Leden: prof.dr. J. Capmany (Universitat Politecnica de Valencia)
prof.dr. L. Pavesi (Università degli Studi di Trento)
prof.dr. B.J. Offrein (Universiteit Twente)
dr.ir. S. Stuijk

Het onderzoek of ontwerp dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

A catalogue record is available from the TU/e Library.

Title: Semiconductor Optical Amplifier-based Photonic Integrated Deep Neural Networks

Author: Bin Shi

Eindhoven University of Technology, 2022.

ISBN: 978-90-386-5545-1

NUR: 959

Keywords: Artificial Neural Networks / Photonic Integrated Circuits / Semiconductor Optical Amplifier/ Optical Computing

This research is financially supported by the Netherlands Organization of Scientific Research (NWO) under the Zwaartekracht programma, ‘Research Centre for Integrated Nanophotonics’.

Copyright © 2022 by Bin Shi

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means without the prior written consent of the author.

書山有路勤為徑
學海無涯苦作舟

Persistence is the path approaching to the summit of profession;
Challenge is the boat to explore the endless ocean of knowledge.

Abstract

Artificial neural networks (ANN) have been employed in a plethora of applications, specifically for complicated problems the conventional modelling cannot easily solve, such as feature extraction, image classification, time series prediction and system optimization. However, the extraction of adequate information from databases remains still a challenge as it requires enormous power and processing time, if artificial neural network models run on conventional electronics. For this reason, electronics community is now busy developing non-von Neumann computing architectures to enable information processing with an energy efficiency down to a few pJ per operation. However, the processing speed of ANN is still constrained to few GHz.

Neuromorphic photonics is an emerging research field that develops an alternative approach to electronics with the attempt to set a milestone in increasing computing speed and decreasing energy efficiency using photons, instead of electrons. The advantages of the parallel nature of light are now being exploited via photonic integrated neural networks based on coherent field summation and/or wavelength-division multiplexing (WDM) optical power addition based schemes to facilitate the massive parallel computation in ANN. However, crosstalk, noise accumulation, insertion losses and low dynamic range prevent further scalability. Moreover, recent works on neuromorphic photonics has been relying on hybrid integrated linear and nonlinear functions, requiring extra chip-to-chip connections, or involving electro-optical conversions, preventing further scalability of photonic neural networks. This thesis demonstrates the implementation and investigation of scalable photonic integrated deep neural networks which utilizes semiconductor optical amplifier (SOA) technology.

InP integration technology has enabled sophisticated photonics circuit design, co-integrating both passive and active elements. SOA is a commonly used component in optical switches to reach high scalability by compensating the waveguide loss. We adopted this technology for neuromorphic photonics to enable high weighting dynamic range and provide optical signal gain on chip. To this end, a photonic feed-forward neural network is demonstrated via an 8×8 Indium Phosphide (InP) cross-connect chip, where up to 8 linear neurons (weighted addition circuits) are integrated on-chip, based on SOA and array waveguide grating (AWG) technologies. The exploitation of these technologies as neural network is evaluated by implementing a trained 3-layer photonic deep neural network to solve the Iris flower classification problem. Prediction accuracy of 85.8% is achieved, with respect to the 95% accuracy obtained via a computer. A comprehensive analysis of the error evolution in our system reveals that the electrical/optical and digital-to-analog conversions dominate the error contribution, bringing down the overall accuracy, and suggests that an all-optical approach is preferable for future neuromorphic computing hardware design.

In order to study the influence of the optical crosstalk from AWG and the waveguide loss, as well as to identify margins for further scalability per layer and energy savings, the same neural network structure is numerically simulated. The weighted addition operation is analysed as a function of the optical crosstalk and the number of inputs (colours). The optical crosstalk results in high errors for the linear operation, while a larger number of channels helps improve performance. The Iris flower classification is simulated using the same neural network structure as in the experiments. The analysis of the prediction error as a function of the optical crosstalk per layer suggests that the first layer performance has a higher impact on the final accuracy.

To enhance the scalability and computing speed of photonic integrated ANN, instead of utilizing a hybrid or electro-optical conversion scheme, an all-optical neuron has been demonstrated to enable tens of GHz processing speed. Specifically, a wavelength converter is exploited as nonlinear function, based on cross-gain modulation in SOA, and co-integrated with the previously verified SOA- and AWG-based linear unit on a single chip. The impact of fully monolithically integrated linear and nonlinear functions on the neuron output is investigated as a function of the number of synapses/neuron and data rate. The monolithically integrated neuron shows better accuracy than the corresponding hybrid device at the same data rate, which is built with photonic integrated linear unit and discrete optical nonlinear function. This all-optical neuron scheme is then used to simulate a 2-layer photonic deep neural network with 64 inputs, 64 neurons at the hidden layer and 10 neurons at the output layer for handwritten digit classification, showing an 89.5% best-case accuracy at 10 GS/s.

Furthermore, the energy consumption is studied for the synaptic operation, considering the full end-to-end system, which includes the transceivers, the optical neural network and the electrical control part. This investigation shows that when the number of synapses/neuron is >18, the energy per operation is <20 pJ, six times higher than when considering only the optical neural network. In addition, the computation speed of this 2-layer all-optical neural network (AONN) system is 47 TMAC/s, 2.5 times faster than state-of-the-art GPUs, while the energy efficiency is 12 pJ/MAC, 2 times better. This result underlines the importance of scaling photonic integrated neural networks on chip.

For the scalability investigation, a model of the noise evolution is proposed for cascaded SOA-based AONNs, including an extended analysis of the noise from the SOAs and wavelength converter. The model is derived from the noise figure estimation and the use of the small-signal method. We demonstrate and simulate the scalability of SOA-based AONN by tuning the optical signal to noise ratio (OSNR) at the neuron input. Experimental results have been used to calibrate the noise evolution model and thereafter to investigate the depth of the AONN. We believe this model will be a valid reference also for the photonic community working on optical switching based on a series of linear and nonlinear SOAs. The exploitation of multiple and interconnected all-optical neurons (AONs) is emulated to be capable of establishing a 12-input/neuron 12-neuron/layer arbitrary layer number all-optical neural network, with a final normalized root mean square error < 0.1, with optimized input signal power at -20 dBm per channel, for a channel spacing of 100 GHz and a gain bandwidth of 32 nm, which can be provided by the recent on-chip SOA.

Moreover, a free-space CNN system is implemented to evaluate its ultimate performance for reduce the computing complexity and identify the role of possible photonic integration. Specifically, the photonic integrated InP cross-connect chip is explored to implement a small-scale convolutional neural network (CNN), by utilizing time-, space- and wavelength-multiplexing schemes, including the cycling properties of the AWG. With 10 Gbaud/s input modulation speed, a 64 SOA filter stage is capable of processing 16 WDM inputs from two free space ranges for a speed of 10.24 TMAC/s, which is 8 times faster than state-of-the-art cross-bar architectures, while it is possible to be reconfigured even in the nanosecond range. The AWG-SOA-AWG structure paves the way to ultra-fast photonic integrated CNN accelerators.

Furthermore, an alternative approach for all-optical neural networks with coherent schemes is discussed to enable dense photonic integration, for which some novel designs are envisioned and their realization is considered on a higher refractive index contrast platform. The mentioned photonic neural networks are benchmarked with respect to the state-of-the-art neuromorphic engines, including electronics and photonics, in terms of computing power, footprint and energy efficiency. The presented SOA-based neural network on InP generic technology does not win in footprint, but it can still guarantee the same performance as conventional electronics with 7 nm CMOS process. In addition, a perspective neuromorphic computing structure is developed, employing a 3-dimensional (3D) integration scheme and exploiting best-in-class available technologies. Here, InP technology, and specifically InP nano-photonics, still play a key role, together with ultra-low loss propagation platforms and novel phase changing materials. With its estimated compactness, ultra-high efficiency and lossless interconnectivity, a 3D neuromorphic photonic engine is foreseen to allow peta-scale computation and ultra-low latency and is expected to shape the future of neuromorphic computing.

Acronyms

AC	Alternating Current
ADC	Analog-to-Digital Convertor
AI	Artificial Intelligent
ANN	Artificial Neural Networks
AO	All-Optical
AO-1L	All-Optical with single layer
AODNN	All-Optical Deep Neural Networks
AON	All-optical Neuron
AONN	All-Optical Neural Network
AO-TL	All-Optical with multiple layers
APD	Avalanche Photodetector
ASE	Amplified Spontaneous Emission
AWG	Array Waveguide Grating
B2B	Back-to-Back
CNN	Convolutional Neural Networks
CNN	Convolutional Neural Networks
CMOS	Complementary Metal-oxide-semiconductor
CW	Continuous Wave
DAC	Digital-to-Analog Converter
DC	Direct Current
DL	Deep Learning
DMUX	De-multiplexing
DNN	Deep Neural Networks
DPO	Digital Phosphor Oscilloscope

DSP	Digital Signal Processor
EDFA	Erbium Doped Fiber Amplifier
EO	Eletro-Optical
FPGA	Field-Programmable Gate Array
FSR	Free Space Range
FWM	Four-Wave Mixing
GMAC	Giga-MAC
GPU	Graphics Processing Unit
IMOS	InP Membrane On Silicon
IP	Internet Protocol
IPDR	Input Power Dynamic Range
IQ	In-phase and Quadrature
ITRS	International Technology Roadmap for Semiconductors
ITO	Indium-tin-oxide
ITU	International Telecommunication Union standard
MAC	Multiplied Accumulation
MMAC	Mega-MAC
MMI	Multi-mode Interferometer
MNIST	Modified National Institute of Standards and Technology
MQW	Multi-Quantum Wells
MRR	Micro-ring Resonator
MUX	Multiplexing
MZI	Mach-Zehnder Interferometer
NL	Nonlinear
NLAF	Nonlinear Activation Function
NL-TF	Nonlinear Transfer Function
NRMSE	Normalized Root Mean Square Error
NRZ	None-Return-to-Zero
OBPF	Optical Bandpass Filter
OE	Opto-Electronic

ONN	Optical Neural Networks
OOK	On-Off Keying
OSA	Optical Spectrum Analyzer
OSNR	Optical Signal to Noise Ratio
PC	Personal Computer
PCM	Phase Change Material
PCNN	Photonic Convolutional Neural Networks
PD	Photodetector
PDNN	Photonic Deep Neural Networks
PIC	Photonic Integrated Chip
PNN	Photonic Neural Networks
PRBS	Pseudo Random Bit Sequence
QCSE	Quantum Confined Stark Effect
RC	Resistor-Capacitor
RGB	Red Green Blue
SCH	Separate Confinement Heterostructure
SOA	Semiconductor Optical Amplifier
SOA-WC	SOA-based Wavelength Converter
SOI	Silicon-on-Insulator
SSE	Spontaneous Source Emission
TE	Transverse Electric modes
TMAC	Tera-MAC
TO	Thermo-Optical
TPA	Two-photonic Absorption
TPU	Tensor Processing Unit
VS	Vector Selection
WDM	Wavelength-Division Multiplexing
WM	Weight matrix Multiplication
Xbar	Cross-bar

Contents

Abstract vii

Acronyms xi

Chapter 1 Introduction **1**

1.1. The Demand of Computing Power 1

1.2. Neuromorphic Electronics 3

1.3. Neuromorphic Photonics 5

1.3.1. CMOS compatible Silicon photonic AI chips 7

1.3.2. InP based photonic AI chips 8

1.4. Challenges on Integrated Photonic Neural Networks 8

1.5. Novel Contributions of this Dissertation 10

1.6. Organization of the Dissertation 11

Chapter 2 Photonic Synaptic Operation **13**

2.1. WDM SOA-based Linear Neuron 13

2.2. PDNN with an InP Cross-connect 16

2.2.1. Weight tuning and calibration 19

2.2.2. On-chip weighted addition and neuron output optimization 21

2.3. Photonic DNN for Image Classification 24

2.4. Error Contribution Analysis 28

2.5. Two-Layer All-Optical Neural Network 32

2.5.1. Experimental setup 32

2.5.2. Experimental results 34

2.6. Conclusion 37

Appendix 2.A. Simulated Investigation on Crosstalk and Path-loss Impact. 39

2.A.1. Implementation of weight calibration and weighted addition 42

2.A.2. Image classification via a three-layer PDNN 45

2.A.3. Energy consumption versus physical layer impairments 50

2.A.4. Conclusion 53

Chapter 3 SOA-based Photonic Integrated All-Optical Deep Neural Network 54

3.1. All-Optical Deep Neural Network 54

3.2. SOA-based Integrated All-optical neuron 58

3.2.1. Experimental setup 59

3.2.2. Integrated SOA-based non-linear function 60

3.2.3. All-optical monolithically integrated neuron 63

3.3. MNIST Classification with an SOA-based AONN 65

3.4. System Energy Consumption Analysis 68

3.5. Conclusion 76

Appendix 3.A Preliminary assessment of Monolithically Integrated AON 77

3A.1 Experimental setup 78

3A.2 Experimental results 79

Chapter 4 Noise Evolution and Scalability Investigation of SOA-based AONN 83

4.1. Noise Modelling of All-optical Neuron 83

4.2. Noise Evolution Emulation 87

4.2.1. SOA characterisation 90

4.2.2. The error evolution of AON 91

4.2.3. The scalability of AON 95

4.3. Conclusion 98

Appendix 4 99

4.A. Back-to-Back measurement 99

4.B. Conversion Efficiency 100

4.C. NRMSE versus Input Noise 102

4.D. Input Noise Suppression Condition 103

4.E. All-optical Neuron Operation with Multi-level Inputs 104

Chapter 5 Photonic Convolutional Neural Network 109

5.1. Convolutional Neural Network 109

5.2. Optical CNN with a 4-f System 111
5.2.1. Experimental setup 111
5.2.2. Results 114
5.3. Photonic CNN with SOA-based Cross-connect 117
5.3.1. Structure and operation of PCNN 117
5.3.2. AWG-SOA-Coupler filter structure 118
5.3.3. AWG-SOA-AWG filter structure 120
5.3.4. The performance of SOA-based PCNN 122
5.4. Conclusion 126
Chapter 6 All-optical Photonic Neural Networks: alternative 2D and 3D approaches 128
6.1. All-optical neural network on IMOS platform 128
6.1.1. MZIs with novel crossing-based thermo-optical phase shifter 129
6.1.2. All-optical nonlinear function on IMOS 131
6.1.3. Integrated coherent all-optical neural network 132
6.2. Benchmark with the State-of-the-arts 133
6.3. 3D Stacked PNN 136
6.3.1. Top layer—Synaptic operation 136
6.3.2. Middle layer—Routing layer 138
6.3.3. Bottom layer—Nonlinear function 140
6.3.4. 3D hybrid integration approach in perspective 141
6.3.5. Projected performance of the 3D stacked PNN 142
6.4. Conclusion 144
Chapter 7 Conclusions and Outlook 146
7.1. Conclusions 146
7.2. Outlook 150
Bibliography 153
Publication List 169
Acknowledgements 173
Curriculum Vitae 178

Chapter 1

Introduction

This chapter introduces the challenges of computing and data processing in terms of scalability, speed and power consumption. To address these challenges, Non-von Neumann structure is utilized to increase the computing speed and the energy efficiency. Neuromorphic computing is recognized to be one of the most efficient ways to carry out fast computation. *Neuromorphic photonics* is an emerging research field that exploits the intrinsic parallelism of light to increase computing speed and energy efficiency, comparing to electronic. This dissertation will focus on photonic artificial neural networks (ANN) which, owing to their versatile nature, can find application in data features extraction and image classification, in time series prediction and system optimization problems, to mention few of them. Photonic neural networks have been so far implemented considering both coherent and wavelength division multiplexing (WDM) approaches. After an introduction to neuromorphic computing and its state-of-the-art, these two architectural approaches will be introduced in this chapter to motivate my work on all-optical neural networks and architecture, as well as to highlight the innovative aspects of this thesis.

1.1. The Demand of Computing Power

We are creating every day nearly half of the information of the human genome on this planet, i.e. about 2.5 Exabytes of data, and every year we keep generating more than what we have created in the past. The total annual global IP traffic is estimated to reach 4.8 Zettabyte per year by 2022 [1]. This exponential increase of data generation in the world is leading to new paradigms in data processing, analytics, exploration and utilization, and is now heavily supported by artificial intelligence. While conventional computing technologies and architectures have

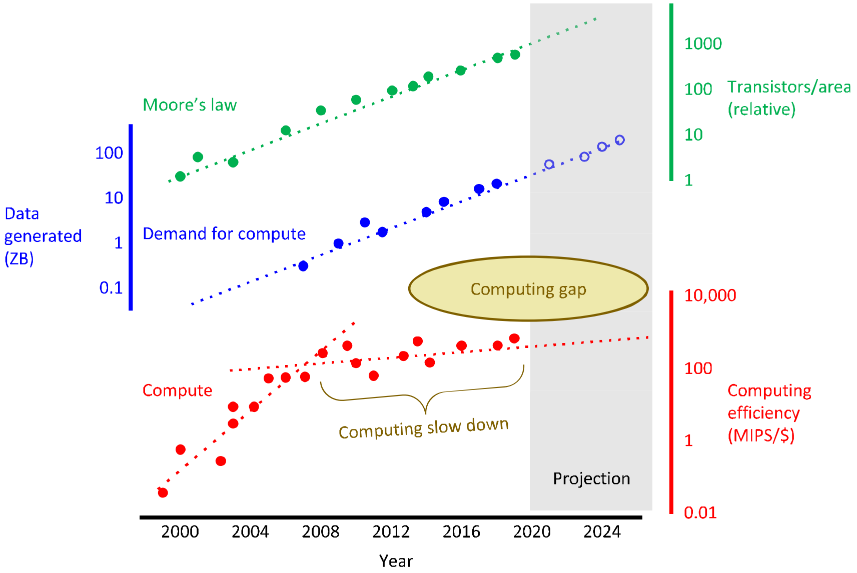


Fig. 1.1 Highlight of the increased demand for compute power, and the 'computing gap' between the generated data and computing power since 2008. [2],[3].

been successful for decades, they are now limiting advances and sustainability in a computer-hungry world.

Device integration in microprocessor chips has been making a steady progress at the pace of Moore's law: the number of transistors per chip has been doubling every two years on average during the last four decades. But Moore's law has been coming to an end: while it is actually still true that we are able to double the number of transistors per chip every year, the clock frequency of processors has levelled off after 2008, when Dennard's scaling law appeared to have broken down. In fact, while Dennard anticipated that both voltage and current would have scaled proportionally with feature size, or similarly that the power would have scaled proportionally with the area, leading to a doubling of performance per watt about every 2 years, actually at smaller sizes, current leakage causes the chip to heat up, preventing further increasing of the clock frequencies. Multiple processors have then helped sustain a steady growth of performance gains by parallel computation, but at the expense of energy consumption. Very soon, however, this new technological trend has faced another issue: the throughput limitation governed by Amdahl's law. In fact, in a multi-core setting, only fractions of the integrated circuit can actually be active at any given point in time without violating power constraints, which means that no effective major gains in performance have been reached since then (see Fig. 1.1), creating the so-called "computational gap" between the amount of data generated and the actual available resources to compute them [2],[3]. This urges us to reinvent our core technologies from the fundamental level through to the architectural level to finally deliver an abundance of computational power.

There is a strong motivation in the microelectronics community behind exploring novel technology frontiers in the “post-Moore era [4]”. The International Technology Roadmap for Semiconductors (ITRS) has already emphasized that the trend for increased performance via “miniaturization” will continue anyhow (*More Moore* approach), eventually trading performance against power, but sustained by the incorporation of new materials and new transistor concepts. At the same time, a functional diversification of semiconductor-based devices will be carried out, where non-digital functionalities will be added (*More than Moore* approach) to facilitate the interconnection with the external world and to power up the overall system, to complement the digital processing and the storage functionalities. This approach will contribute further to the miniaturization of electronic systems, even not at the same rate as before. The co-integration of different technologies to achieve the desired performance gain sees, among other technologies, integrated photonics playing a major role within this roadmap [5]. Alternative solutions are also sought to extend the CMOS roadmap, where nanotechnology is exploited to replace conventional planar MOSFETs [6] and supplement 3D stacked nanophotonic deployments [7]. In this case, we talk about the *Beyond Moore* strategy, which also includes purpose-built processors that accelerate application-specific tasks (such as the Tensor Processing Units from Google, Graphics Processing Units, FPGAs, ASICs, etc.) [8]. These accelerators rely on a very large number of smaller processors where workloads can be broken down and parallelized to run computational demanding applications. However, in microelectronics, the achievable data throughput is eventually limited by the same electrical interconnections, as skin effect, dielectric loss, and wiring density exacerbate power dissipation [9]. This, in turn, forces most of the hardware to stay in idle mode, waiting for the data to be fetched by the memory, to minimize the memory–processor communication energy costs. There is no doubt that any revolutionary chip technology that supplants conventional silicon electronics will certainly have to be interconnect- and memory-centric [10].

1.2. Neuromorphic Electronics

With the continuous advances in microelectronics, supercomputers – the core of the information processing – are now able to execute around hundreds of Peta FLOPS/s (floating point operations/s) [11], but at the impractical cost of about tens of millions of watts. Our brain, in comparison, is able to perform about the same order of operations but consuming only 20W [12]. Designing future hardware circuitry by getting inspiration from brain connectivity promises to offer a real opportunity to overcome the limitations of conventional electronics. New computing paradigms of non-von Neumann architectures have, therefore, begun to unfold, attracting renewed interests during the last decade and leading to the development of a plethora of machines based on novel computing architectures [2], such as neuromorphic or other biologically inspired architectures.

With the diminishing returns on Moore's law and the rise of deep learning (DL) to prominence in 2012, the computing industry has started to rapidly move from a programming- to a learning-era [13]. Spurred by the digital energy efficiency wall and following the neuroscience-focused approach, which aims to replicate fundamentals of biological neural circuits in dedicated hardware, important advances has been made in neuromorphic and DL accelerator ASICs. Some notable examples are IBM's TrueNorth [14], SpiNNaker [15], Intel's Loihi [16], Neurogrid [17], and HiCANN [18]. These have been built to bring the power efficiency down to few pJ (picoJoule) per MAC (Multiply-ACcumulate) operation. These large neuromorphic machines are based on the spiking architectural model: being more complex, these models are still not fully understood, unlike the more advanced Deep Learning models. The rich DL model portfolio can be indeed utilized in digital GPU and TPU engines, as well as in the constantly growing number of emerging artificial neural network (ANN)-based analogue electronic AI chipsets.

Deep-learning-focused approaches, on the other hand, begin with an end-application in mind and aim to construct hardware that efficiently realizes a solution, while eliminating as much of the complexity of biological neural networks as possible. Deep learning has attracted much attention because it is particularly good at learning from unstructured data by using artificial layered neural networks, which makes it potentially very useful for real world applications. Computational architectures based on the interconnectivity of multiple neurons are called artificial neural networks. Deep neural networks are the quintessential deep learning models. These are feedforward models, in the sense that information flows through layers of artificial neurons in one unique direction, from the input toward the output, where the outputs of one layer are fed to the next layer as inputs, and are able to run inference, prior to being trained. In Fig. 1.2, a comparison between a biological neuron and an artificial neuron is shown. The same neuron is then embedded with many others for the construction of a feed-forward artificial neural network.

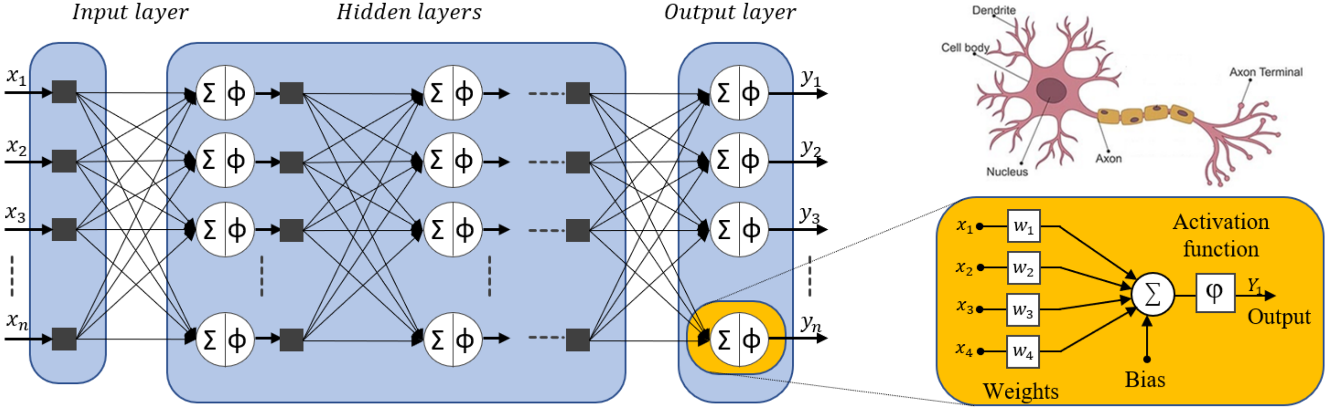


Fig. 1.2 General scheme of a deep neural network made of layers of artificial neurons. The basic model of an artificial neuron (yellow inset) is made of two distinct parts: a linear part (Σ) and a nonlinear part (Φ).

An artificial *neuron* represents the basic operation unit in a neural network. In our work, we refer to the McCulloch-Pitts neuron model [20], which is commonly implemented in an artificial neural network (ANN), like a deep neural network. Then the operation executed by a neuron is modelled as $y = \varphi(\sum W_i x_i + b)$, where φ is the activation function, x_i is the i-th element of the input vector, w_i is the weight factor for the input value x_i and b is the bias. We call the linear term $\sum W_i x_i$ weighted addition. For a layer of M interconnected neurons, the output of these neurons can be expressed in vector form: $\boldsymbol{y} = \varphi(\boldsymbol{W} \cdot \boldsymbol{x} + \boldsymbol{b})$, where $\boldsymbol{x}$ is an input vector with N elements, $\boldsymbol{W}$ is the $N \times M$ weight matrix, $\boldsymbol{b}$ is a bias vector with M elements and $\boldsymbol{y}$ is a vector made of M outputs.

Among the most powerful DL hardware, we name the GPU-based DL accelerators hardware favoured due to their high compute and memory density as well as an established hardware and software ecosystems [21],[22], allowing at the same time for footprint-energy efficiency improvements by ∼7 orders of magnitude on average [23],[24]. In parallel to the digital approach, there is a constantly growing number of emerging artificial neural network (ANN)-based analogue electronic artificial intelligence (AI) chipsets that, following the technology trends in neuromorphic computing, tend to collocate processing and memory to minimize the memory–processor communication energy costs. The most exciting emerging AI hardware architectures, designed to avoid digital bottlenecks, are the analogue crossbar approaches [25]–[27], since they achieve parallelism, in-memory computing, and analogue computing at the same time. Mythic’s architecture [28], for example, can yield high accuracy inference applications within a remarkable energy efficiency of just 0.5 pJ/MAC. These advances follow all plethora of extensively investigated memristive devices based on a variety of physical processes that go from phase changing to spin torque transfer, which have yielded a number of interesting approaches for high-density storage and computing [29]-[31].

Even if the implementation of neuromorphic approach is visibly bringing to quite some outstanding record energy efficiencies and computation speeds, we see that neuromorphic electronics is already struggling offering the desired data throughput at the neuron level: artificial neural networks rely on dense interconnectivity between neurons, but closely spaced wires experience bandwidth-distance trade-offs due to resistor-capacitor (RC) parasitic effects, with current machines hardly exceeding GHz clock frequencies, so that further scaling in computational speed and throughput will result in side effects as a huge energy consumption [32],[33]. Neuromorphic processing for high-bandwidth applications requires about GHz operation per neuron, which calls for a fundamentally different technology approach [34]. A future-proof solution that could dominate this landscape for many years should obviously rely on the best-performing and top-efficient technology and architectural mixture that can support the complete DL learning model portfolio.

1.3. Neuromorphic Photonics

Neuromorphic photonics rises as a new research field [35], which aims to transfer the well-known high-bandwidth and low-energy interconnect credentials of photonic circuitry to the area of neuromorphic platforms. In contrast to electronics, there is negligible energy overhead for moving light encoded information within the photonic processor, which enables unprecedented circuit interconnectivity and speed. Moreover, photonic engines are bit-rate agnostic, offering the right credentials toward bit-rate-transparent operation that can relax the delicate trade-off between speed and power consumption. On top of that, photonic integration technology has reached now a maturity level where high-performance sophisticated integrated circuits are made available [36],[37]. Only the combination of the complementary advantages of photonics and electronics and their synergic co-design will enable processing systems with high efficiency, high interconnectivity, and extremely high bandwidth for the postulation of the new field of neuromorphic photonics at the nexus between photonics and neural network processing models. Breakthrough proof-of-concept experimental photonic AI platforms [38]-[40] have recently started appearing, initially exploiting the maturity of the CMOS silicon photonics industry and shaping a potential path toward integrating multiple photonic neurons on the same silicon chip. Building on these architectures, neuromorphic photonics has come to the fore stark making headlines through a few start-up companies [41]-[44] and raising expectations for orders of magnitude higher energy and size efficiencies compared to state-of-the-art electronic AI platforms [27],[45]-[47]. It is of paramount importance to review and categorize the most notable examples of neuromorphic photonic demonstrations to be able to understand platform and architecture limitations and foresee long-term developments. While also the neuroscience focused approach has been applied to photonics [35],[48]-[51], the works in this dissertation will concentrate on deep learning photonic integrated approaches and will emphasize at the main strategies and architectural approaches for linear neurons, as the synaptic operation is the most computational expensive.

Two main architectural approaches are used to realize linear neurons and connectivity between those when moving from a layer to the next one: the *coherent and noncoherent approaches,* shown in Figs. 1.3a–f. Coherent layouts. [39] rely

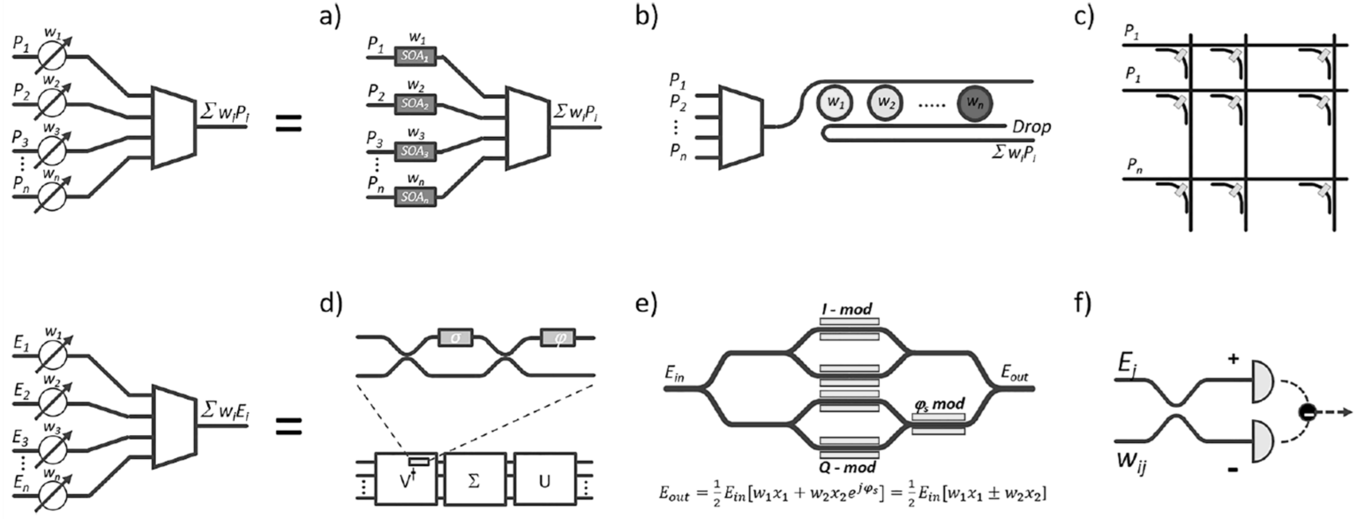


Fig. 1.3 (Top) Linear photonic neuron implementations based on WDM optical power addition: (a) optical neuron for a cross-connect neural layer, (b) a bank of micro-ring resonators for the broadcast and select (B&S) layer scheme, and (c) the crossbar scheme for the parallel computing using PCM loaded micro-bends. (Bottom) Linear photonic neuron implementations based on coherent electric field summation: (d) the optical interference unit with single stage based on two phase shifters MZI, (e) the IQ modulator scheme on InP, and (f) the summation using the concept of the quantum photoelectric multiplier.

on the use of interferometric arrangements that require just a single wavelength for their optical input signal. These layouts can provide addition or subtraction while the weighted signals are still in the optical domain via constructive or destructive interference of the optical beams, respectively, allowing in this way the representation of signed values via encoding at the optical carrier signal phase. On the other hand, non-coherent configurations [35] utilize multiple wavelength channels as their input signals relying on wavelength selective filtering elements for the weighting function and on optical power addition at the photodetector. This requires sign information to be also encoded in the wavelength domain and the final summation to be carried out at the output of a balanced photodetection scheme, where the optical powers of wavelengths carrying the positive and the negative values, respectively, are converted into electrical currents and subsequently subtracted. Finally, recently a first demonstration of combining both coherent and non-coherent approaches has been proposed for multi-functional optical neural network, leading to an increase of data throughput and enable different modes for multi-neuron, convolution and fully connected network operations.

1.3.1. CMOS compatible Silicon photonic AI chips

Shen et al. have recently proposed a coherent approach (Fig. 1.3*d*) using a Mach-Zehnder Interferometer (MZI)-based optical interference unit (OIU) for matrix multiplication combined with software-implemented saturable absorbers to form two-layer feedforward neural networks on silicon on insulator (SOI) [39]. Though this is a promising approach and also extendable to quantum application, the presence of multiple MZI stages for implementing a single weight increases phase noise accumulation, reducing extinction ratio and preventing scalability. Also, the weight is set via a thermo-optical mechanism, which increases power consumption. An optical neural network accelerator, based on time-multiplexing and coherent (homodyne) detection, has been proposed [52], which promises to be scalable to large networks without any error propagation issue (Fig. 1.3*f*).

Many-mode ONN operation has been demonstrated in a free space system using spatial light modulators [54], but a faster operation is anticipated by employing silicon photonic integrated chips. With the addition of the wavelength domain, the micro-ring-resonator (MRR)-based weighting bank facilitates scalability with easy implementation of neurons and interconnections [55], based on wavelength division multiplexing (WDM) optical power addition (Fig. 1.3*b*), showing a computational speed of sub-TMAC/s (Tera MAC/s), a computing density of few TMAC/s/mm^2, and an energy efficiency of already 0.52 pJ/ MAC, similarly to the top electronic AI chip [27]. While this weighting scheme has been demonstrated to achieve up to 6 bit precision [54], this is obtained at the cost of complicated calibration schemes. In all these cases, while the routing happens via the SOI low-loss waveguides, the nonlinear functions are realized via software, off-chip, or via the use of a photodiode (PD) balanced scheme. However, an all-optical (AO) implementation of the neural network calls for complete removal of E/O conversions.

1.3.2. InP based photonic AI chips

With respect to the previous implementations on silicon platforms, indium phosphide (InP) material platform has the advantage to allow the co-integration of active and passive components without a loss in performance. The integration of active elements provides gain compensation for on-chip loss as well as nonlinear function on-chip, which opens to scalability. Excitable lasers based on ring-laser [57] and two-section DFB laser [58], shows premising integrated nonlinear implementation for neuromorphic computing with sensitive thresholding, triggered with input power as low as -6 dBm. Based on the in-phase and quadrature (IQ) modulator scheme, a coherent optical linear unit is demonstrated (Fig. 1.3*e*) for a computational speed of 0.32 TMACs/s and an energy efficiency of 1.5 pJ/MAC [55]. This architecture is exploited for MNIST digit recognition obtaining an average accuracy as high as 97.24%, when combined with the empirical transfer function of the MZI-semiconductor optical amplifier (SOA) optical nonlinear function scheme [56]. This dissertation is aiming to extend the research field of AI chip on InP platform, with the all-optical neural networks with the WDM inputs, base-on array waveguide grating (AWG) and the SOA technology. The linear synaptic operation with integrated SOA-base cross-connect chip, as shown in Fig. 1.3*a*, with the incoherent approach, will be demonstrated in Chapter 2 and monolithically co-integration with SOA-based nonlinear function is shown in Chapter 3.

1.4. Challenges on Integrated Photonic Neural Networks

The recent emerging Neuromorphic Photonics leverages the implementation of photonic deep neural networks to enable high-throughput and lower power consumption computing with the nature of light, as mentioned in section 1.3. To establish a reliable photonic neural network, one need to address the challenges on the photonic implementations.

- **Optical matrix multiplication.** The mathematical calculation of matrix multiplication needs to be carried out in the optical domain. This is at the moment done with the approaches mentioned in section 1.3, which calls for a scalable photonic matrix multiplication unit with high connectivity, high-dynamic range and ease of calibration.

- **Optical nonlinear function.** Computing nonlinear function in the optical domain area asked to provide ultra-fast processing of the data in 10s to 100GHz speed. However, so far photonic integrated linear and nonlinear functions have been demonstrated to rely on hybrid integration schemes, hindering the realization of a scalable photonic neural network. All-optical neural network implementation, based on all-optical neurons, is expected to offer a route to scalability.

- **Scalability.** The scalability of the PDNN defines the maximum size of a neural network, the size of a DNN defines the number of parameters which can be set and therefore the size of the problem that we can actually solve. Therefore, it is desirebla that the interconnectivity of the neurons in a network, and the depth of the signal propagation in the forward direction are big enough for solving a certain cluster of problem applications, e.g., image classification, as well as easy to reprogram for an ad-hoc tailored depth, depending on the application case.

- **Noise suppression.** To enable the scaling of the PDNN, one must handle the noise accumulation that is present after the optical signal processing. A photonic implementation which supresses noise is required for PDNN, in order to achieve non-degraded signal processing, and therefore guarantee a high accuracy.

- **Low power consumption.** The power consumption of the data processing in the photonic processor should outperform the electronics in terms of energy per operation. Hence, high throughput and low power weighting need to be considered to increase the energy efficiency.

- **Compact integration.** The integration of the photonic neural network on chip should allow for a compact footprint, in order to integrate even more neurons on a chip and enable large-scale PDNN.

- **Fast reconfiguration.** The reconfiguration of the weighting matrix in the PDNN is usually to be considered in low speed since the inference of the testing data is applied to a network which is trained in advance. However, the acceleration of training of the neural network demands fast tunning of the weight matrix, in the µs range for 10 GHz data processing for 10k input samples, which is still a missing feature in state-of-the-art DNN implemented in the photonic domain and calls for further study in the near-future.

1.5. Novel Contributions of this Dissertation

This dissertation presents the research work on SOA-based linear operation unit with a photonic integrated cross-connect chip and the co-integration of all-optical nonlinear function to form a monolithically integrated all-optical neuron. The performance of the all-optical DNN is analysed and simulated, to identify computing power and efficiency, and a noise analytical model is developed for the WDM chip operation for anticipating scalability. The implementation of photonic convolutional neural network is also discussed to achieve ultra-fast computation. The main novel contributions of the dissertation are listed as follows:

- The SOA-based cross-connect photonic integrated chip (PIC) is proposed to be exploited as linear matrix multiplication unit, utilizing the combination of AWG and SOA technologies. The WDM scheme with the AWGs enables high-throughput interconnection, while the use SOAs provide gain for the scaling and high dynamic range for the weight tuning. The weight calibration per neuron is performed with a normalized root mean square error smaller than 0.08 and a best-case dynamic range of 27 dB. A 3-layer PDNN with off-chip nonlinear functions is demonstrated to solve a pattern classification problem, resulting in 85.8% classification accuracy. The comprehensive analysis of the error evolution in the system reveals that the electrical/optical conversions dominate the error contribution. This suggests that an all-optical approach is preferable for future neuromorphic computing hardware design.
- The first monolithically photonic integrated all-optical neuron including an SOA-based weighted addition function and a wavelength converter with a tunable laser as nonlinear function is demonstrated. Performance analysis in terms of input data rate and signal power are comprehensively assessed. The monolithically integrated neuron shows better accuracy than the corresponding hybrid device at 10 GS/s input. The energy consumption for synaptic operation is also analysed, considering the full end-to-end system, which includes the transceivers, the optical neural network and the electrical control part. This investigation shows that when the number of synapses/neuron is >18, the energy per operation is <20 pJ (6 times higher than when

considering only the optical engine). The results indicate that an all-optical layer scaling on-chip results in a reduced power consumption per MAC operation.

- A novel noise model is proposed for investigating the signal degradation on the signal processing after cascades of SOAs in the SOA-based all-optical neuron, for WDM operation. The model is validated via experimental results. Both experiments and simulations reveal that the all-optical neuron, with multi-wavelength to single wavelength conversion as non-linear function, is able to compress noise from the optical inputs and the amplifiers after a certain number of layers. This suggests that the use of SOA-based all-optical neuron (AON) with wavelength conversion may allow building feed-forward neural networks with arbitrary depth. The result shows that the on-chip integrated AONs is capable to create 7-input 7-neuron/layer PDNN with arbitrary depth with guaranteed error less than 0.1 when operating at input power levels of -20 dBm/channel and with a 6 dB input dynamic range.
- A free-space convolutional neural network (CNN) system is implemented based on an optical correlator which realizes Fourier transform, in order to evaluate its ultimate performance and identify the role of possible photonic integration. The photonic integrated InP cross-connect chip is discussed to implement a small-scale CNN, by exploiting 4-dimension of parallelism in the space, wavelength, FSR and cyclic wavelength domains. The integrated cross-connect chip, consisting of 8 AWG-SOA-AWG filter structures with 8 SOAs per filter, is capable to process 10.24 TMAC/s with an energy efficiency of 0.26 pJ/MAC for PCNN implementation.

1.6. Organization of the Dissertation

This dissertation includes the collective works from the implementation of SOA-based linear unit of deep neural network towards monolithically integrated AONN and paves the way to all-optical SOA-based photonic integrated noise stable arbitrary cascaded deep neural networks.

In Chapter 2, the implementation of SOA-based neural network on the linear unit will be explained. The operation of the weighting SOA is demonstrated in detail, including the linear weighted addition structure and the weight calibration. An optical cross-connect on-chip is exploited as one layer of neurons and, with reconfigurations of the weight SOAs, as a three-layer deep neural network, to demonstrate a pattern classification problem both experimentally and via simulations.

In Chapter 3, the architecture of an all-optical neural network with SOAs in linear and nonlinear operation region is proposed. The co-integration of SOA-based synaptic operations is presented, which includes a combination of SOAs and AWGs for the linear part, and of a wavelength converter, based on cross-gain modulation as the nonlinear operation.

In Chapter 4, the noise evolution and scalability of the proposed SOA-based AON with noise modeling are studied along three dimensions: multi-level, multi-wavelength and multi-layers and the depth scalability is investigated.

In Chapter 5, the operation of the convolutional neural network (CNN) is explained and demonstrated in free space with 4-f optical correlator system. A photonic integrated version of a CNN is also proposed via the integrated cross-connect, exploiting the multiplexing dimensions of space, wavelength, FSR, and cyclic wavelength.

In Chapter 6, the proposed all-optical SOA-based photonic deep neural network is benchmarked with respect to the state-of-the-art electric and photonic DNN implementations. We further discuss the possibility to miniaturize AONN by moving to a 3-dimension integration technology.

In Chapter 7, the conclusion of the research works is drawn, and the outlook of the future works is discussed.

Chapter 2

Photonic Synaptic Operation

In this chapter, the implementation of an SOA-based linear neural network will be presented and explained. The operation of the SOA-based weighting is demonstrated in section 2.1, which includes also the linear addition circuitry and the weight calibration. An optical cross-connect, integrated on-chip, is explored as one layer of linear neurons in section 2.2. By reconfiguring the matrix of weight-SOAs, a three-layer deep neural network is experimentally demonstrated to solve a pattern classification problem (section 2.3). An in-depth error analysis is made to understand the major sources of error in section 2.4. Finally, in section 2.5 a preliminary investigation of a two-layer all-optical neural network, realized with the fully integrated linear operation, combined to discrete non-linear optical components, is carried out with promising results.

2.1. WDM SOA-based Linear Neuron

As mentioned in section 1.2, a *neuron* represents the basic operation unit in a neural network. In this work, the artificial neuron refers to the McCulloch-Pitts neuron model [20], which is commonly implemented in an artificial neural network (ANN), like a deep neural network (DNN). Then the operation executed by a neuron is modeled as $y = \varphi(\sum W_i x_i + b)$, where φ is the activation function, x_i is the i^{th} element of the input vector, w_i is the weight factor for the input value x_i and b is the bias. The *weighted addition* in a neuron is given by $\sum W_i x_i$. For a layer of M interconnected neurons, the output of these neurons can be expressed in vector form: $\boldsymbol{y} = \varphi(\boldsymbol{W} \cdot \boldsymbol{x} + \boldsymbol{b})$, where $\boldsymbol{x}$ is an input vector with N elements, $\boldsymbol{W}$ is the $N \times M$ weight matrix, $\boldsymbol{b}$ is a bias vector with M elements. The M outputs form the $\boldsymbol{y}$ vector. For a feed-forward DNN, the outputs of one layer are fed to the next layer as inputs.

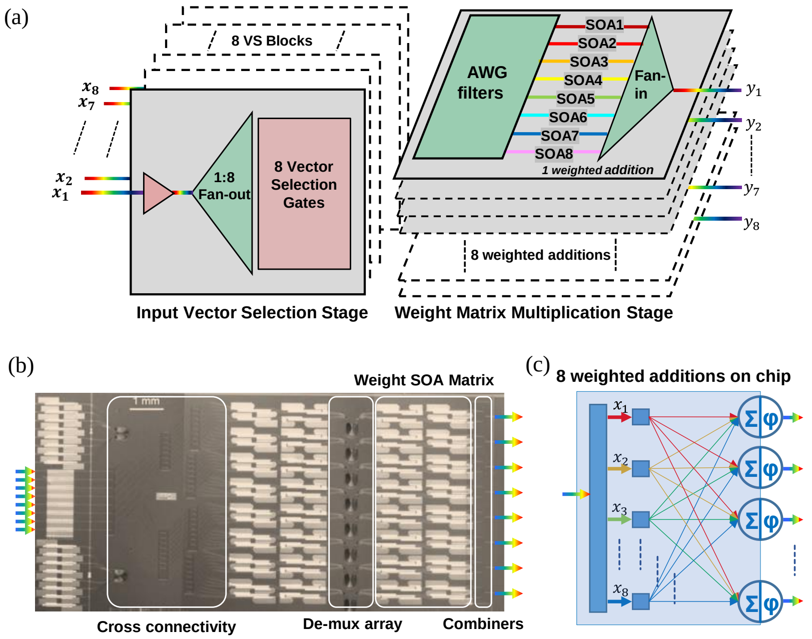


Fig. 2.1 (a) Schematic diagram for the chip which co-integrates 8 weighted addition operation circuits for 8 WDM input vectors and provides 8 WDM outputs. In grey are highlighted the used blocks. (b) Composite microscope image of the fabricated PIC. (c) The functionality implemented on chip for 1 WDM input corresponds to one layer of 8 weighted additions in a DNN, with off-chip activation function.

A cross-connect photonic integrated switch for optical communication application [59] is employed in this paper since it embeds already the connectivity required for feed-forward networks, as described above (see Fig. 2.1). The 8×8 InP SOA-based cross-connect chip is capable of providing space connectivity of up to 8 neurons, and multi-wavelength connectivity of up to 64 channels, eight multiplexed channels (a WDM signal) per input. Here the SOA technology is exploited in combination with the arrayed waveguide grating (AWG) technology

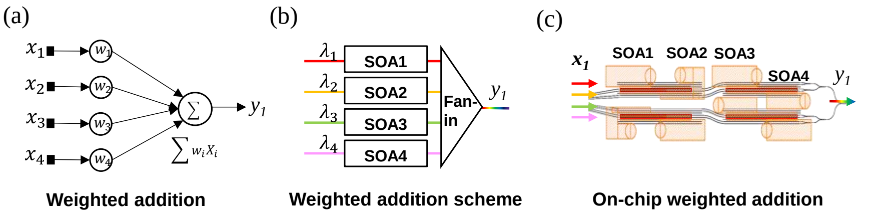


Fig. 2.2 (a) Scheme of the weighted addition within one neuron, (b) components and (c) mask details of the circuitry designed on the integrated InP cross-connect.

for multiple reasons: The optical amplifiers are employed for setting the weight matrix and providing on-chip gain for scalability, while the AWGs are used to filter out the out-of-band noise built up by cascading multiple stages of SOAs, as well as to de-multiplex the input data channels. The scheme in Fig. 2.1*a* includes a pre-amplified input Vector Selection stage (VS, left boxes), which includes a fan-out unit and an input vector selection unit. A Weight Matrix multiplication stage (WM, at the right) is made of AWG-based filters, followed by the weight-SOA matrix and an addition (fan-in) stage.

Fig. 2.1*b* shows a composite image of the PIC used in this paper. This is realized on an active-passive InP substrate. The active regions are used for SOAs and include four InGaAsP/InP quantum wells with optimization for TE polarization. The passive regions for the waveguides, splitters and cyclic AWGs use the same 500 nm-thick confining hetero-structure but without the quantum wells. Both active and passive confinement layers are sandwiched between InP cladding layers. The weight SOA elements are 1mm long and 2 µm wide. Deep-etched, high sidewall verticality, 1.5 µm-wide waveguides are used for the low-radius curved-waveguides, the power splitters and the curved waveguides within the arrayed waveguide gratings. Deep-etched multi-mode-interference devices are used for power splitters and combiners. Shallow-etched, low-loss, 2 µm-wide ridge-waveguides are used for most of the circuit layout. SOAs and shuffle network waveguide crossings are also implemented as shallow waveguides.

The chip broadcasts the WDM signals to eight neurons belonging to the same layer and de-multiplexes it via an array of eight AWGs. Fig. 2.2*a* shows the scheme of the weighted addition implementation of a neuron and the mask details of the neuron layout on-chip. Each neuron consists of eight (different channels) inputs, eight SOAs (one per channel) and an 8:1 MMI (multimode interferometer) based optical combiner. In particular, the AWGs right after the input vector selection allow for a reduced out-of-band accumulated noise due to the pre-amplification stages, in order to increase the weight resolution. By applying different currents to the SOAs (shown in Fig. 2.2*b*), the multiple multiplications between the input channels and the trained weight-SOA matrix are performed. The output signals are then summed up to finalize the weighted addition operation. The thresholding function φ is not co-integrated in this first demonstration (see Fig. 2.1*c* and Fig. 2.2*c*). In this chapter, the weighted addition of 4 input channels per neuron for a total number of four (best performing) neurons per layer is demonstrated.

The weighted addition is formulated mathematically, including the use of SOA-based weights and of WDM data inputs. The WDM signal at the input of the chip includes N multiplexed channels, and each channel λ_i represents one element x_i of the input vector. For a given input power x_i, and when the internal losses are neglected, the expression of the output of the multiplication of an SOA weight by the input power can be expressed as [60]:

$$x_o = x_i \exp[h_i(I_i)], \tag{2.1}$$

where h_i is the gain integrated over the length of the SOA for *path*-i, where the data is transmitted over the channel λ_i, for a certain injection current I_i. The SOA is assumed to be a broadband component so that the gain function is constant with the wavelength over a wide band. Once each channel is multiplied by the gain $\exp[h_i(I_i)]$, all results are summed again into a WDM signal via an on-chip optical combiner. The signal amplitude detected at a photodetector (PD), placed right at the chip fiber output, is:

$$v = \sum (R \cdot Z_0 / v_\pi) \cdot x_i \cdot \exp[h_i(I_i)]\,, \tag{2.2}$$

where R is the signal detection response, assumed to be constant for dense WDM signals, Z_0 is the PD characteristic impedance and v_π the voltage for a π phase shift [61].

The bias and the activation function are implemented off-line via a computer for this first demonstration. The value of the bias $\boldsymbol{b}$ is added to the data after detection. The activation function is chosen to be a hyperbolic tangent function $\varphi(v) = \tanh(\alpha v)$, where α is the slope of the *tanh* function. The power transfer function of the SOA is very similar to the positive part of the *tanh* function [60], so that an activation function realized with nonlinear SOAs is also possible within the InP platform. A procedure for fine optimization of the weights to the nominal trained weights, of the *bias* and of the slope of the *tanh* function is carried out and demonstrated in section 2.2.2 to improve the output accuracy of the neurons.

2.2. PDNN with an InP Cross-connect

The experimental set-up used in this paper is shown in Fig. 2.3. Four tunable lasers (Santec, ECL-210) are operated at wavelengths λ_1=1539.83 nm, λ_2=1543.10 nm, λ_3=1546.26 nm and λ_4=1549.06 nm, each one modulated with an electrical pattern generated by an arbitrary waveform generator (AWG, Tektronix, AWG7122B). A channel separation of about 3.2 nm is chosen to match the on-chip AWG characteristics and optimal output channel transmission. In this section, 1-bit modulation is used for demonstrating weight calibration and weighted addition operation. The patterns are amplified by an electrical amplifier (SHF 100APP) to drive the 4 V_{pp} optical amplitude modulators (SUMITOMO, T.MXH-1.5-10PD-ADC) that generate the optical data signals. The experimental setup

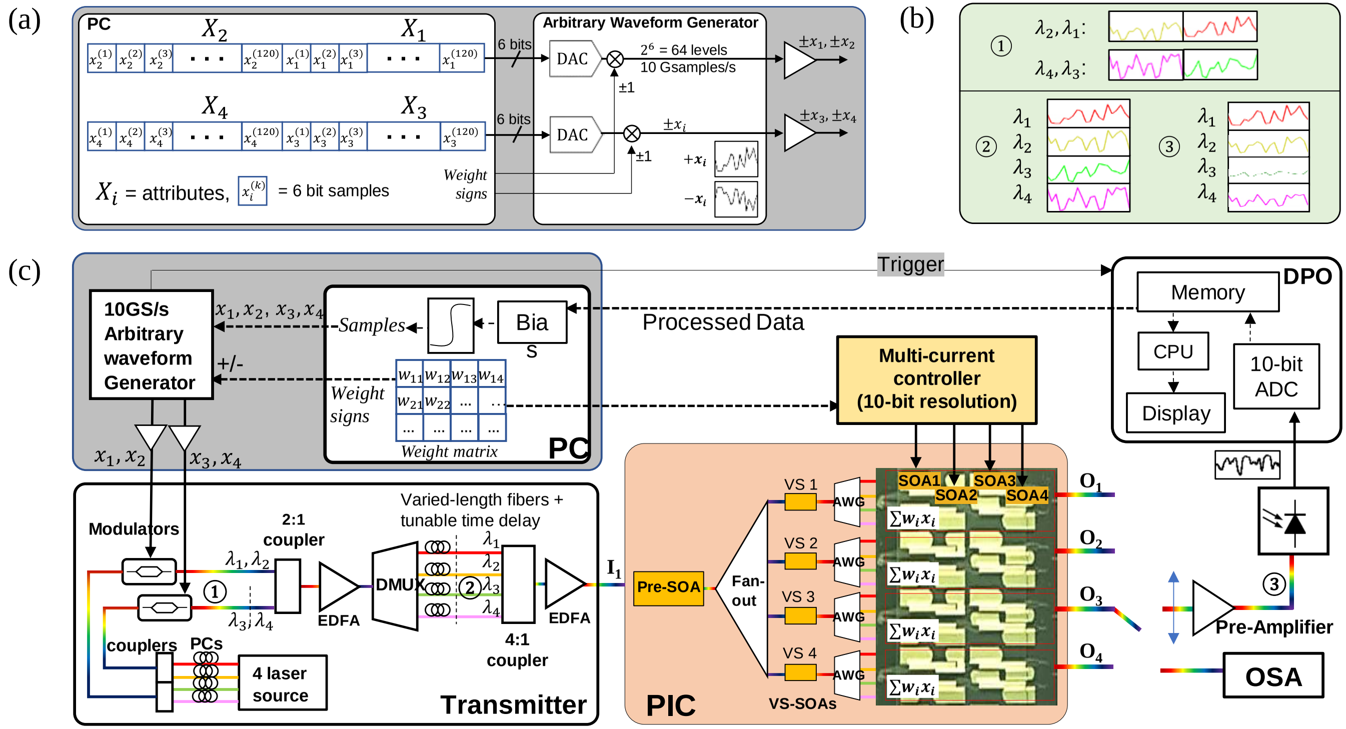

(a)
PC
X_2
X_1
X_4
X_3
$x_2^{(1)}$ $x_2^{(2)}$ $x_2^{(3)}$ ... $x_2^{(120)}$ $x_1^{(1)}$ $x_1^{(2)}$ $x_1^{(3)}$ ... $x_1^{(120)}$
$x_4^{(1)}$ $x_4^{(2)}$ $x_4^{(3)}$ ... $x_4^{(120)}$ $x_3^{(1)}$ $x_3^{(2)}$ $x_3^{(3)}$ ... $x_3^{(120)}$
X_i = attributes, $x_i^{(k)}$ = 6 bit samples
Weight signs
6 bits
6 bits
Arbitrary Waveform Generator
DAC
DAC
2^6 = 64 levels
10 Gsamples/s
±1
±1
$\pm x_i$
$+x_i$
$-x_i$
$\pm x_1, \pm x_2$
$\pm x_3, \pm x_4$
(b)
① λ_2, λ_1:
λ_4, λ_3:
② λ_1 λ_2 λ_3 λ_4
③ λ_1 λ_2 λ_3 λ_4
(c)
10GS/s Arbitrary waveform Generator
x_1, x_2, x_3, x_4
+/-
Samples
Bias
Weight signs
w_{11} w_{12} w_{13} w_{14}
w_{21} w_{22}
Weight matrix
PC
x_1, x_2
x_3, x_4
Trigger
Processed Data
Multi-current controller (10-bit resolution)
DPO
Memory
CPU
Display
10-bit ADC
Modulators
2:1 coupler
λ_1, λ_2
λ_3, λ_4
①
EDFA
DMUX
Varied-length fibers + tunable time delay
λ_1 λ_2 λ_3 λ_4
②
4:1 coupler
EDFA
couplers
PCs
4 laser source
Transmitter
I_1
Pre-SOA
Fan-out
VS 1
VS 2
VS 3
VS 4
AWG
SOA1 SOA2 SOA3 SOA4
$\sum w_i x_i$
VS-SOAs
PIC
O_1 O_2 O_3 O_4
③
Pre-Amplifier
OSA

Fig. 2.3 (a) Coding scheme used in the experiment. (b) Time traces for the different wavelengths at point 1, 2 and 3 of the setup. (c)Experimental setup. The chip circuitry is shown in the grey box, where the four weighted addition circuitries is shown as part of the chip. PCs: polarization controllers; DMUX: de-multiplexer; Pre-SOA: semiconductor optical preamplifier; VS: input vector selector gate; AWG: Array waveguide grating; EDFA: erbium doped fiber amplifier; PD: photodetector; OSA: optical spectrum analyser; DPO: Digital phosphor oscilloscope; PIC: Photonic integrated circuit. Tunable time delay indicated with dashed lines.

allows the use of only two modulators (instead of four) for implementing the data encoding into 4 different wavelengths. This is possible by combining the use of the described coding scheme with the implementation of signal synchronization in the optical domain. The generated signals are then combined to form one unique WDM signal, which is amplified using an erbium doped fiber amplifier (PriTel LNHPFA-22) and de-multiplexed again into four amplified wavelength channels. Polarization control is performed on each channel to maximize SOA gain on chip. The data in the four different wavelengths are then synchronized using varied-length fibers and tunable time delays before being combined again and being input at the PIC port 1 via a lensed fiber. Up to 4 outputs are accessed with a lensed fiber which is scanned at the chip output and couples the output signals to an optical spectrum analyzer (OSA, HP7145B) for spectra inspection, or to a pre-amplified AC coupled PD (DSC-R402APD) and a Digital Phosphor Oscilloscope (DPO, Tektronix, DSA72004C). Here, the time traces are digitized via a 10-bit ADC and recorded for calculating the normalized root mean square error (NRMSE). These time traces are input to a computer where both the non-linear function and an optimization procedure are implemented. The neuron outputs are then sent back to the pattern generator for implementing deeper neural networks.

In the PIC scheme (Fig. 2.3, pink box), one neuron operation requires the biasing of up to 6 SOAs: 1 SOA used as pre-amplifier (orange Pre-SOA), one SOA used to select the input vector (orange VS_i) and 4 SOAs acting as weights (SOA1, SOA2, SOA3 and SOA4 in orange). The operation of a layer of a total number of four weighted additions requires the biasing a total number of 21 SOAs: 1 pre-amplifier SOA, four SOA for selecting the input vectors and 16 SOA acting as weights. The 16 weight-SOAs are biased with different (weight) currents controlled by a multi-current controller (Thorlabs MLC8200-8CG), in order to assign the gain value which acts as a *weight factor* to the corresponding input data. The DACs for the weights and for the input data are separated in different devices. In particular, the Arbitrary Waveform Generator includes two 8-bit DACs and provides the electrical analog signals to be encoded in the optical CW laser signals at the transmitter side. Instead, the control currents used to set the weight-SOAs are set by the multiple current sources with a resolution of 10-bits (50 µA resolution).

The implemented coding scheme is shown in Fig. 2.3*a*. It is conceived to allow input data bit-precision adaptability from 1 up to 6 bits. In Fig. 2.3*a* the coding scheme is depicted for the example of 6-bit precision. Four different attributes, $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, $\boldsymbol{X_3}$ and $\boldsymbol{X_4}$, are coded via a computer as two trains of 6-bit precision samples, x_i, which are min-max normalized [62] to the range [-1,1]. This procedure of normalization does not impact the final outcome *y*, as long as corresponding new values for the slope of the non-linear function, *α'*, and for the bias, *b'*, are set. These two signals are then sent to two inputs of the 10 GS/s Arbitrary Waveform Generator (AWG): the attributes $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$ are serially sent to input 1 of the AWG, while the attributes $\boldsymbol{X_3}$ and $\boldsymbol{X_4}$ are serially sent to input 2 of the same AWG. There, they are converted into analog signals via two 8-bit DACs in the range [-V, V], where V is the reference voltage at the DACs.

In this experiment only positive weights are set, as weighting via an SOA is inherently positive. Therefore, in order to perform all combinations of multiplications between positive inputs, x_i, and positive/negative weights, w_i, we encode ad-hoc the input data for every neuron output so that the multiplication with a negative weight is equivalent to multiplying that same weight, but positive, by the input signal, now inverted. Specifically, the output of each DAC in the AWG will provide for the analog signal of the original data, x_i, or the inverted analog signal of the original data, $-x_i$, depending on if the sign of the weight, communicated via computer to the AWG, is positive or negative, respectively. Alternatively, negative numbers are possible via the complementary modulation scheme suggested in [63], that though requires to double the number of WDM inputs to the photonic integrated chip.

In the next sections, we calibrate the weight-SOA gain (section 2.2.1), perform weighted addition of 4 input channels per neuron and optimize it via three feedback loops (section 2.2.2).

2.2.1. Weight tuning and calibration

In this work, the SOA gain is exploited as weight matrix element w_i, since an SOA can deliver a wide dynamic range when employing both the absorption and the amplification regimes. However, the weight-SOA gain curve is not a fully linear function of the injected current. Calibration of the gain-current curve is needed in order to correctly reconfigurate the weight. In this first experiment, we modulate an optical input signal with 10 Gb/s pseudo random bit sequence (PRBS) and we record the optical peak power curves as a function of the current injection at an OSA. For this first prototype, the fabricated AWG has an optical crosstalk of -19 dB [59], meaning that each optical path will influence the gain control of the other channel paths. Therefore, the used calibrated tuning scheme considers an average operation condition for the optical power when the optical crosstalk is maximum and when it is minimum (the impact of the optical crosstalk and path losses is investigated via a series of simulations on VPI and reported in Appendix 2.1). The optical crosstalk is expected to be the highest when all weight-SOAs

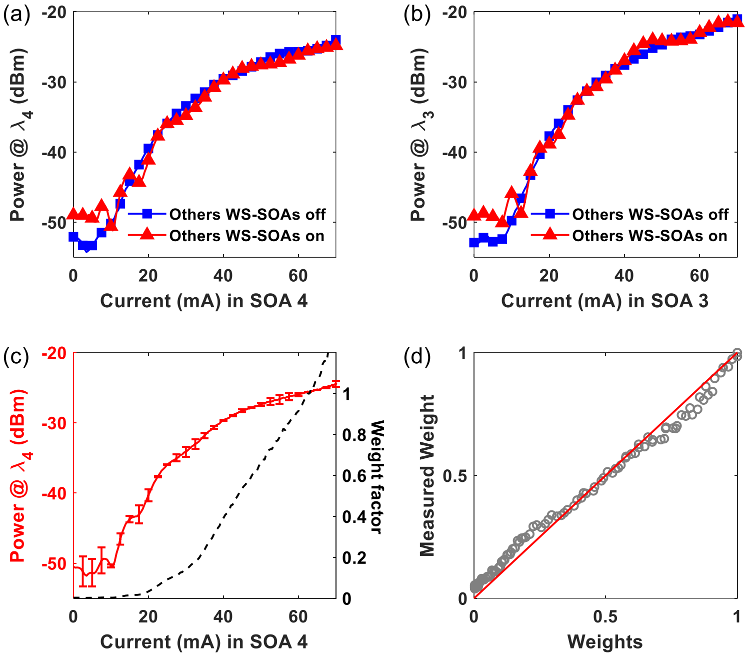


Fig. 2.4 Measured optical power when tuning the injection current at weight-SOA4 (a) and weight-SOA3 (b), when the other weight-SOAs are off (blue curve) and when the other weight-SOAs are biased at 70mA (red curve). (c) Mean curve (solid line) from measurements in (a) with error bars, and the weight factor curve versus current (dashed line) with power at -25.5 dBm taken as reference point. (d) Correlation plot of the recorded weights versus the set weights after calibration.

are switched on and biased with a maximum injection current of 70 mA (Fig. 2.4*a*, red triangles). On the other side, the crosstalk on the channel under operation will be minimum if all the other SOAs are switched off (Fig. 2.4*a*, blue squares). Fig. 2.4*a* presents the optical peak power curves measured when increasing the injection current of the controlled weight SOA4 from 0 to 70 mA, with all the other weight-SOAs are on and off, shown in red triangles and blue squares, respectively. These curves indicate an optical power tunable range from -49.5 dBm to -25.0 dBm, which enables a dynamic range of 24.5 dB. The same measurements are carried out for the other channels as well. Fig. 2.4*b* plots the output power as a function of the control current for the SOA3, for the same conditions as in Fig. 2.4*a*: This offers a dynamic range of 2.5 dB wider. The difference in optical power between the 2 SOAs is due to path dependent losses. Higher dynamic ranges are possible when crosstalk is improved. The curve oscillations present in both Fig. 2.4*a* and *b* might be due to interference between the crosstalk signal and the signal in the desired path. The on-state crosstalk from other paths is already visible from the offset between the blue curves and the red curves at 0 mA. By averaging the power of the two curves in Fig. 2.4*a*, the optical power control curve is obtained as a function of the current with error bars. A maximum error of about 5% (0.6 dB) in the exploited range from 15 mA to 70 mA is recorded, but this will be far reduced by using optimization feedback loops as explained in the following section 2.2.2. The weight factor curve is generated by setting a reference point at weight '1' for a reference output power. For full operation of all the channels, this reference point is defined by the lowest of all the weight-SOA output powers measured at the maximum current of 70 mA that, for this chip, corresponds to -25.5 dBm output power. The weight factor curve versus current is then plotted in Fig. 2.4*c* (dashed line). In the obtained weight factor curve, one can observe that there are 3 quasi-linear operation regimes: i) for a current value within the range 0-15 mA, the weight factor is negligible; ii) for a current value in the range 15-35 mA, the weight factor can be tuned with a resolution of 0.01 per mA; iii) for current values within the range 35-70 mA, the weight can be tuned with a three-time lower resolution of 0.03 per mA. In this setup, the minimum control current step is 0.05 mA. After obtaining the weight calibration curves, from the plot in Fig. 2.4*d*, a linear relation between the set and the measured weights is obtained, meaning that the weight-SOA curve is well calibrated with the current: a maximum error of 0.045 is reported.

2.2.2. On-chip weighted addition and neuron output optimization

After the calibration of all the four SOA-weights per neuron, the weighted addition operation in a neuron is carried out. All four optical input channels are now modulated with 10 Gb/s PRBS. The NRMSE between measured and calculated two-channel (λ_3 and λ_4) weighted addition at the neuron output is reported in Fig. 2.5*a* as a function of the assigned weight to channel λ_4. A fixed weight factor '1' is assigned to channel λ_3. An NRMSE below 0.08 is recorded when the pre-amplifier is set to operate in transparent condition (solid line in Fig. 2.5*a*). A higher pre-amplification provided by the on-chip pre-SOA and the input selection-SOA (4 dB net added gain) increases the in-band noise and therefore results in a worst NRMSE of up to 0.11 (dashed line in Fig. 2.5*a*). However, in a newly designed chip, these pre-amplifiers are not required. The 4-channel weighted addition is demonstrated with randomly set currents/weights for all the four SOAs of one neuron over all the weight factor range [64]. As an example, the measured (blue) and calculated (red) weighted addition time traces are shown for a part of the bit sequence, and a reading error of 0.06 (Fig. 2.5*b*, top) and 0.04 (Fig. 2.5*b*, bottom) is calculated. In accordance with the different current operation regime in Fig.

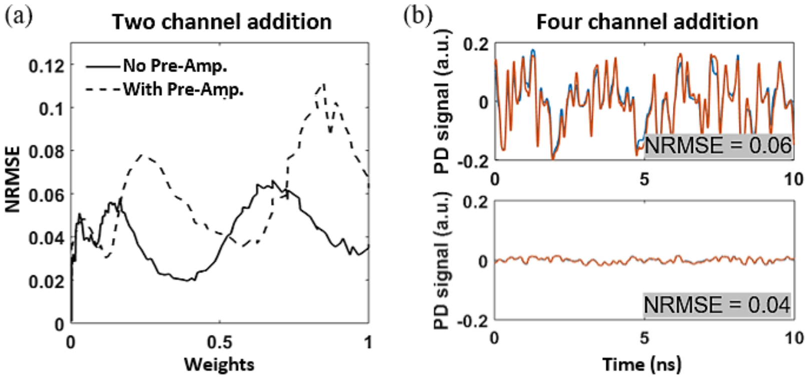


Fig. 2.5 (a) Measured weighted addition reading error from two-channels (λ_3 and λ_4) as a function of the set weight on λ_4, when net on-chip pre-amplification is zero (solid line) and 4 dB (dashed line). (b) Measured 4 channel weighted addition time traces (blue) and the calculated traces (red) when randomly setting current of four weight control SOAs of one neuron.

2.4*c*, the weight set accuracy is higher for lower weight factor values. Since the receiver at the chip output is an AC coupled Avalanche Photo-Detector (APD), filtering out the DC bias from the detected signal, the signals in Fig. 2.5*b*, measured at the output of the APD, range from negative to positive values.

Three electronic control feedback loops of the neural network parameters are implemented as shown in Fig. 2.6, in order to optimize the output accuracy of a neuron: fine tuning of the control current at the SOAs (loop-*i*), tuning of the bias (loop-*ii*) and of the non-linear function shape (loop-*iii*) [65]. These 3 loops are executed independently and for one neuron, while keeping fixed the other parameters.

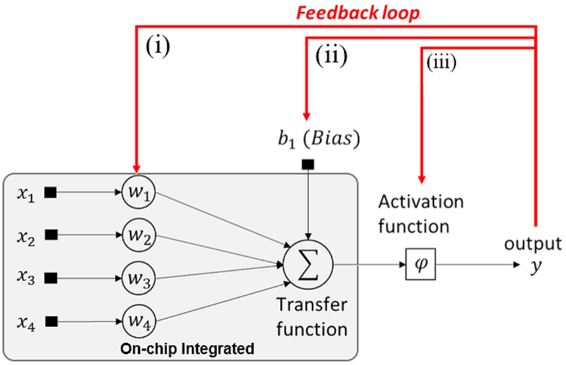


Fig. 2.6 Optimization loops for improving the NRMSE of a neuron (after the activation function). Error optimization (i) of the individual weights at the neuron, (ii) of the offset to the trained bias and (iii) of the trained non-linear function shape.

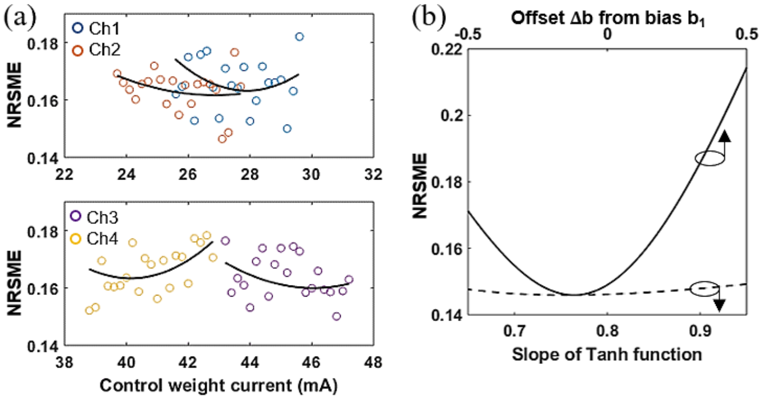


Fig. 2.7 Error optimization of the individual weights at a neuron (a), of the offset to the trained bias (b, solid line) and of the non-linear function shape (b, dashed line).

Fig. 2.7*a* shows the outcome of the optimization procedure at the weighted addition operation on chip which utilizes 4 channel input with a given weight matrix (loop-*i*). The neuron output error variation is recorded (after the implemented activation function) when tuning the control current near the calibrated weight w_i. The weight of each channel is tuned from w_i-0.02 to w_i+0.02 with a weighting step of 0.005. The solid lines are parabolic curve fitted out of the measured points, showing the (local minima) optimal points. After optimization of the on-chip set weights, the best NRMSE of the neuron is recorded to be 0.15 (non-optimized NRMSE was 0.2). Furthermore, the bias is tuned by changing an offset Δb to the trained bias and the shape of the *tanh* function is changed by varying the slope α, with respect to the values obtained from the simulation model (loop-*ii* and loop-*iii*). Fig. 2.7*b* plots the calculated error when tuning the bias offset (solid line) and when changing the slope of the hyperbolic tangent function (dashed line). The impact of the offset tuning on the NRMSE is remarkable as it acts on all the data at the same time. On the other side, the NRMSE changes quickly for values of *tanh* function near to zero but slowly for values at the edge of the range [-1, 1], as expected.

2.3. Photonic DNN for Image Classification

In order to evaluate the performance of a deep neural network built via multiple layers on the InP photonic cross-connect, we chose to solve a pattern classification problem. One of the most famous classification problems is the Fisher's Iris flower classification, which can be solved by using a DNN. The Iris database is made of three classes (setosa, versicolor and virginica) of 50 instances each [66]. Per each instance, the Iris flower category is identified by observing four of its attributes: length and width of sepals and petals. For this demonstration, we have executed the training of this DNN via the simulation platform *TensorFlow* [23], where we have used 120 instances as a training database. In order to make use of 4 weighted addition operation circuitries available in the photonic integrated InP cross-connect chip for four different neurons in a layer, a feed-forward network, made of 2 hidden layers with 4 neurons and an output layer with 3 neurons (see Fig. 2.8), is trained on a computer. The dashed boxes in Fig. 2.8 indicate the same chip, which is used for three times in this experiment, as it only includes one layer of linear neurons. The 3-layer network is created layer by layer. For every layer, we assign the weight factor to different inputs on the same chip and record the optical signal from the output of the chip. This is converted into the electrical domain and passed through the nonlinear function on the computer. The results are then fed back to the transmitter side to generate again the optical inputs. For the next two layers, the procedure is equivalent. This 3-layer network is used to demonstrate the concept of an SOA-based photonic neural network. In particular, the Iris flower classification problem has been chosen as a small scale example to analyze the ultimate accuracy provided by the photonic neural network.

The trained weight matrix is mapped to the matrix multiplication on chip. The hyperbolic tangent activation function is implemented off-line via software after

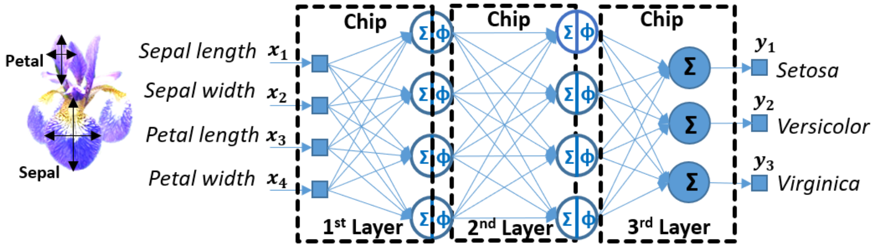


Fig. 2.8 The deep neural network for the Iris flower classification trained on a computer. The dashed boxes indicate where the InP cross-connect chip is employed. Each neuron is indicated as made of the weighted addition part (Σ) and the non-linear function (φ). The latter is out of the box as it is not implemented on chip.

the O/E conversion. The output from the first hidden layer serves as input to the second layer (at the arbitrary waveform generator) and the output from the second hidden layer serves as input to the third (output) layer. Finally, the output of the third layer, after the *SoftMax* transfer function, $P(y = j) = e^{y_j} / \sum_{j=1}^{n} e^{y_j}$, provides the predicted probability of the output samples y belonging to class j.

For this application, the 4 input attributes are encoded into 2^6 levels at a data rate of 10 GSample/s and are used to modulate the four input wavelengths which are amplified and multiplexed to be input at port 1 of the chip. The fine tuning of the weights for output optimization (as described in section 2.2) is applied to each neuron in every layer of the photonic neural network. The expected values of the output signals are calculated as the input data (coming from the previous layer) multiplied by the corresponding trained matrix. The difference between the expected values and the actually recorded values at the output of the chip, after the optimization procedure, represents the real performance of the chip, i.e. it tells how well the trained matrix can be mapped onto the implemented photonic neural network. Both the time traces of the expected values (red line) and the recorded optimized signals (blue line) are plotted in the same graphs in Fig. 2.9 for all the four neurons in the first layer. The recorded signals are similar to the expected ones, where the calculated NRMSE are 0.15, 0.10, 0.08, and 0.10 for neuron 1 to 4 (from top to bottom), respectively.

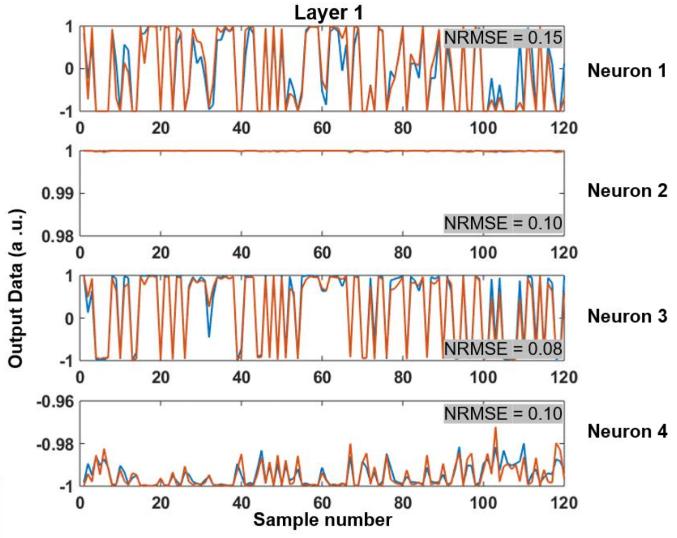


Fig. 2.9 Output data from the 1st hidden layer, with blue line representing the measured data after optimization and red line being the expectation from the input multiplied with the trained matrix.

For the 2nd hidden layer and 3rd layer (the output layer), the same procedure is carried out to optimize the outputs. TABLE 2.1 shows the NRMSE of the optimized signal at the output of the i^{th} layer per neuron, compared to the expectation calculated taking the output from the previous $(i-1)^{th}$ layer and using it as input to the i^{th} layer. TABLE 2.2 shows the percentage of improvements obtained per layer and per neuron: The optimization loops contribute to improving NRMSE over a range from 5% up to even 55%. A better optimization is observed when the output data is lying on the saturation region of the hyperbolic tangent, or in other words when the output data are pushed to the edges of the '±1' range of *tanh* output (see Fig. 2.9, traces from neuron 2 and 4). For the output layer 3, all the improvements are roughly 5%, due to the fact that the optimization procedure is executed only on the linear combination of the neuron outputs of the layer.

Correlation matrices are used to show how accurate is the final prediction made by the photonic neural network. In order to understand the role of the PIC into the changes in final prediction accuracy, this is calculated in three cases: the on chip weighted addition operation is implemented only within layer 1 and the output is fed to the remaining simulated network (Fig. 2.10-A), the on chip weighted add-

TABLE 2.1 NRMSE, comparing outputs to the expectations

Output	Neuron 1	Neuron 2	Neuron 3	Neuron 4	Average
Layer 1	0.15	0.10	0.08	0.10	0.10
Layer 2/1	0.04	0.07	0.03	0.10	0.06
Layer 3/2	0.08	0.07	0.05	-	0.07

TABLE 2.2 Achieved NRMSE improvements by the optimization

Output	Neuron 1	Neuron 2	Neuron 3	Neuron 4	Average
Layer 1	25%	50%	27%	50%	37%
Layer 2	55%	5%	25%	23%	27%
Layer 3[a]	5%	5%	5%	5%	5%

[a] there is no activation function at output layer, the optimization is merely on the weighted addition part on chip.

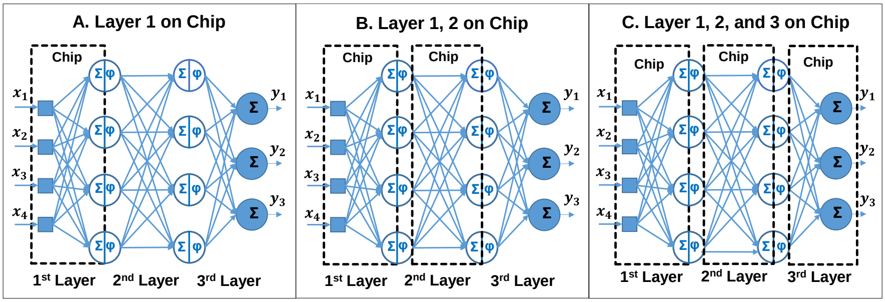


Fig. 2.10 For a deeper understanding of the photonic deep neural network, the photonic weighted addition layer is implemented within (A) the 1st layer only, (B) the 1st and 2nd layer, and (C) within all the 3 layers of the DNN.

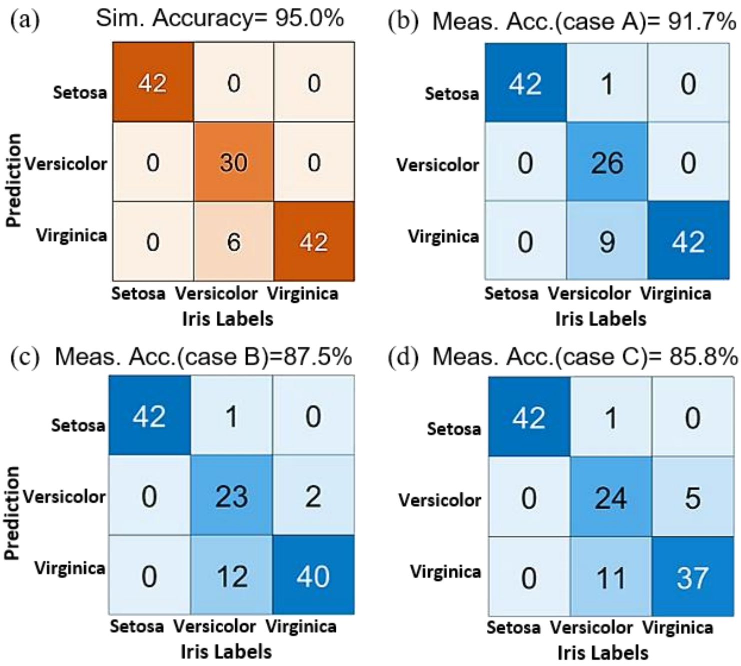


Fig. 2.11 (a) Simulated label prediction of the DNN indicates an accuracy of 95%. (b) Experimental image prediction using the photonic weighted addition within the 1st hidden layer (case A in Fig.2.10). (c) Experimental image prediction using the photonic weighted addition within the 1st and the 2nd hidden layer (case B in Fig.2.10). (d) Experimental image prediction when using the photonic weighted addition within all the 3 layers (case C in Fig.2.10).

ition operation is implemented within layer 1 and layer 2 and the output from layer 2 is fed to the remaining simulated network (Fig. 2.10-B), and finally the on-chip weighted addition operation is implemented within all the three layers (Fig. 2.10-C).

In Fig. 2.11*a* we display the prediction accuracy coming from the simulation on the computer (PC): This is calculated to be 95% as 6 versicolor, out of 36 instances, are not well predicted. After adding the on-chip integrated weighted addition layer, we also calculated the accuracy prediction in case A, B, and C of Fig. 2.10. The prediction accuracy decreases when the number of layers implemented on a photonic neural network increase, changing from 91.7% accuracy (Fig. 2.11*b*, case A) down to 85.8% (Fig. 2.11*d*, case C). This may not be mainly attributed to the on-chip integrated networks: at the layer interfaces, the information is converted several times from the electric into the optical domain, and in the electrical domain again, which introduces additional impairments. A comprehensive error analysis is carried out in the next section to discriminate error contributions.

2.4. Error Contribution Analysis

We analyze all the error contributions to understand any physical limitation of the present implementation. The total error σ_{N_j} measured at the output of every neuron N_j is the summation of the signal impairments induced by different stages σ_k: the distortion at the modulation, at the weighted addition operation within the photonic chip, at the detection path, and the error induced by applying bias and activation function. It can be described as:

$$\sigma_{N_j} = \sigma_{mod} + \sigma_{chip} + \sigma_{detc} + \sigma_{bias} + \sigma_{act}.$$

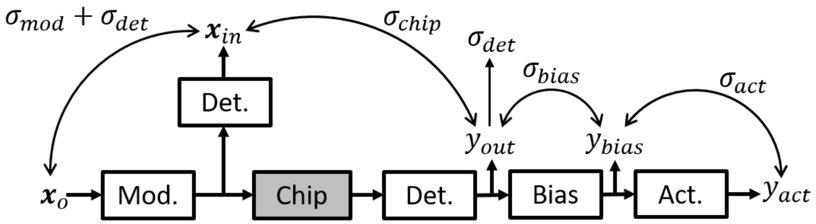


Fig. 2.12 The error propagation through the different stages in a layer. The error contributions from different stages comes from the error differences between adjacent stages.

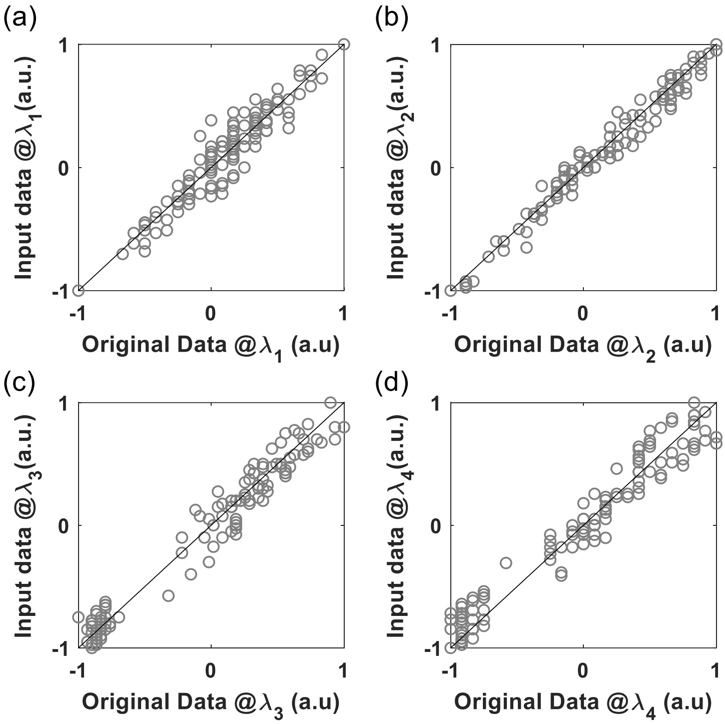


Fig. 2.13 Correlation of the data after modulation at the input channel with respect to the normalized original data, for (a) channel 1, (b) channel 2, (c) channel 3, and (d) channel 4.

The error from modulation, σ_{mod}, includes the error from the distortion induced by the electrical amplification after pattern generator for driving the optical amplitude modulator. Error from the chip, σ_{chip}, quantifies the order of impairments building up within the chip due to crosstalk and chip imperfections. The error from the detection path, σ_{detc}, includes the distortion error from the preamplifier and the noise at the photodetector. The error σ_{bias} from the bias and the error σ_{act} from the activation function are important contribution as they may enhance previous errors.

All these error contributions can be accessed by measuring the signal error at the output of the different stages in a layer as illustrated in Fig. 2.12. The data at the input and at the output port of the chip, at the detector output, after the bias, and after activation function are measured. Then the error contributions σ_k are calculated as error difference between the adjacent stages. The total error is expected to accumulate while passing through the different stages.

The modulation induces optical signal distortion. The data after the modulator are recorded with a 0.5 nm pass bandwidth filter and plotted with respect to the normalized original data input from the computer. Fig. 2.13 shows the correlation of the data at the chip input port and the original data set at the input of the first layer, after normalization. The NRMSE is then calculated and shown for different channels from channel 1 (a) to channel 4 (d). The average of these errors gives the σ_{mod} at the input of the first layer. The modulation errors of all the input channels from λ_1 to λ_4, at the input of the first hidden layer, of the second hidden layer and of the third (output) layer are listed in Table 2.3. The calculated average errors σ_{mod} at the input of the first hidden layer, of the second hidden layer and of the output layer are 0.03, 0.05 and 0.06, respectively.

The detection error is obtained by calculating the maximum error between the detected data and the average data from N times detection. Due to phase noise, the measured error ranges from 0.015 to 0.026. σ_{detc} is set to a maximum fixed value of 0.03 and assumed to be identical for all the wavelengths.

From Fig. 2.14a to c, the error contributions of different stages of the computation from the 1st hidden layer to the output layer are shown. From the results in Fig. 2.14a and b, the activation function shows to be an important factor in the error estimation. The reason why the $tanh$ function induces massive error (see for example for Neuron 1 of Layer 1) is due to the fact that part of the output data from the chip is close to zero, whose error does easily enhance after the activation function. Vice versa, the activation function does not incur in extra losses but, on the contrary, improves the performance (see for example Neuron 4 of Layer 1) when the output data is lying at the edges of [-1, 1] range, since the activation function tends to compress the data. The error contribution from the modulation is the net major contribution to the total final output error, while the matrix multiplication on chip contributes to the error for less than the 0.03, i.e. less than the error due to the detection path. Removing the error from the electro-optical (E/O) conversions would significantly reduce the total error at the output.

After analyzing the error distribution per layer of neurons, we also analyze how every layer contributes to the final performance of the implemented neural

TABLE 2.3 Distortion from Modulation σ_{mod}

Input	Ch 1	Ch 2	Ch 3	Ch 4	Average
Layer 1	0.04	0.02	0.03	0.04	0.03
Layer 2	0.08	0.01	0.11	0.01	0.05
Layer 3	0.05	0.06	0.07	0.06	0.06

network. For estimating the error incurred by one layer σ_{L_m}, the average of the error at the output of every neuron j, σ_{N_j}, is considered:

$$\sigma_{L_m} = \frac{1}{n}\sum_{1}^{n} \sigma_{N_j} \,.$$

Then the average error of the 1st layer, of the 2nd layer and of the output layer is 0.10, 0.06, and 0.07, respectively (empty circles in Fig. 2.14*d*). In addition, we estimate the error accumulation at the output of each layer: this is 0.10, 0.16, and 0.23 at the different layers, respectively (empty triangles in Fig. 2.14*d*). The real distortion of the data from experiments to the simulated trained DNN model on the computer is calculated to be 0.10, 0.15 and 0.20 at the output of the 1st layer, of the 2nd layer and of the output layer, respectively (full circles in Fig. 2.14*d*): The estimated accumulation line follows well the calculated error line. Finally,

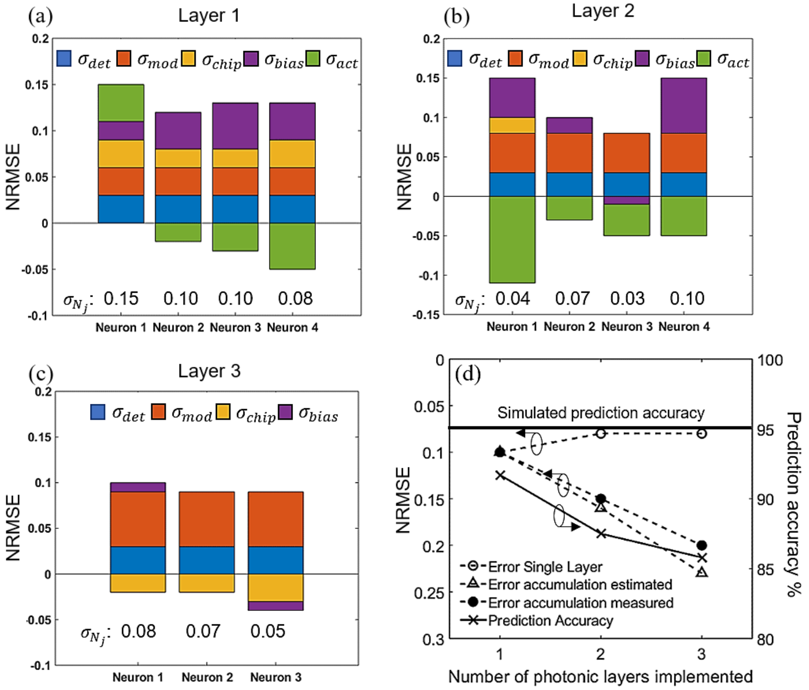


Fig. 2.14 The error analysis from 5 different impairments contributions at (a) Layer 1, (b) Layer 2, and (c) Layer 3. (d) Error evolution versus the number of photonic layers implemented. The empty circles show the error from single layers, the empty triangles plot the estimated error accumulation, the full circles plot the measured error. The cross symbols show the final prediction accuracy. The solid line on top highlights the final prediction accuracy simulated via computer.

the final prediction accuracy of the network is shown via a line of crossings in Fig. 2.14*d*: this matches with the trend of the error evolution from the first to the last layer, as expected. Moreover, while the accumulated error increases up to 0.2, the prediction accuracy is reduced by only 9.2% with respect to the simulated prediction accuracy. This outcome proofs that the photonic on-chip neural network is robust against the impairments added to the signal under processing.

It is important to note that, for the specific case of the Iris flower classification problem, the resolutions offered by the used DACs for the input vectors and for the weights and by the ADC at the receiver side are not expected to impact the ultimate accuracy. On the other hand, these networks are expected to be fault tolerant: An analysis of the use of lower resolution DACs and ADC for the same Iris classification problem reveals that the use of DACs with less than 4-bit precision does not allow to accurately classify the Iris flower. Also, for this extreme case, an ADC of at least 6 bit precision must be used not to degrade the ultimate accuracy.

2.5. Two-Layer All-Optical Neural Network

In this section, an all-optical two-layer feed-forward neural network is investigated and demonstrated. A nonlinear (NL) SOAs is utilized as optical nonlinear activation function, for which the operation of the NL-SOA based wavelength converter is investigated. The on-chip photonic cross-connect, demonstrated in section 2.1 as linear unit, is then all-optically connected to the off-chip optical nonlinear function to process binary sequences through a two-layer all-optical deep neural network (AODNN).

2.5.1. Experimental setup

For tuning the weight factors of the WDM signals, the broadcast-and-weight (B&W) [68] is used as fully tunable network protocol: The protocol utilizes wavelength division multiplexing (WDM) to provide dense connections between nodes, wherein each node i is assigned a unique wavelength λ_i. By multiplexing many connections within each waveguide, a small number of waveguides can carry signals for many more connections. We realize the basis weighted addition operation, by using sophisticate InP photonic integrated chips, as shown in section 2.1. The input data are encoded into different colors and the weights are tuned by controlling the injection current of (weight) SOAs, therefore modifying the amplification of the input data. In this first demonstration, we use up to two weights per neurons. We implement the all-optical interconnection from the 1st

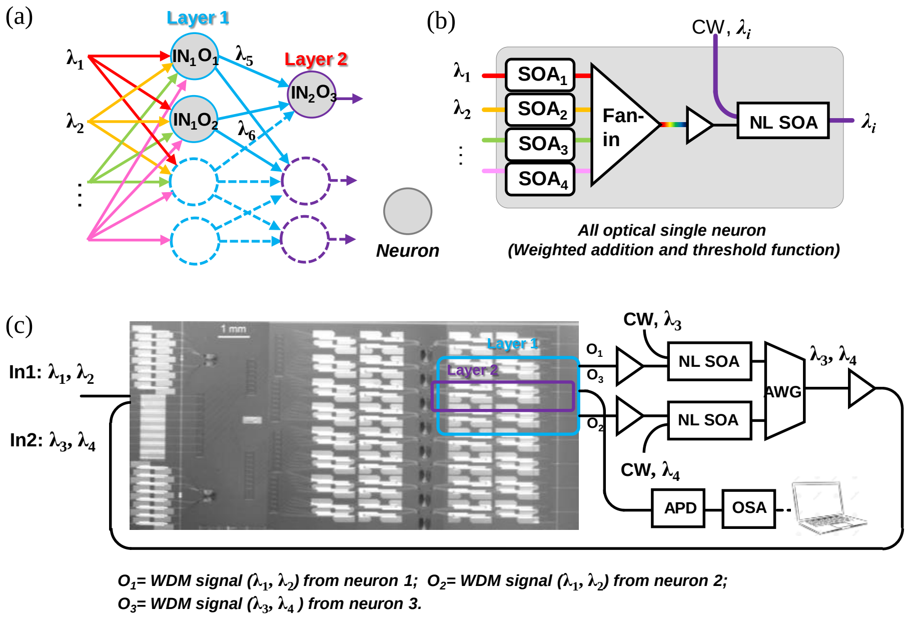

(a)
Layer 1
Layer 2
λ_1
λ_2
IN_1O_1
IN_1O_2
IN_2O_3
λ_5
λ_6
Neuron
(b)
CW, λ_i
λ_1
λ_2
SOA_1
SOA_2
SOA_3
SOA_4
Fan-in
NL SOA
λ_i
All optical single neuron
(Weighted addition and threshold function)
(c)
In1: λ_1, λ_2
In2: λ_3, λ_4
1 mm
Layer 1
Layer 2
O_1
O_3
O_2
CW, λ_3
CW, λ_4
NL SOA
NL SOA
AWG
λ_3, λ_4
APD
OSA
O_1= WDM signal (λ_1, λ_2) from neuron 1; O_2= WDM signal (λ_1, λ_2) from neuron 2;
O_3= WDM signal (λ_3, λ_4) from neuron 3.

Fig. 2.15 (a) Representation of the two-layer all-optical neural network, with gray circles are used neuron for this first demonstration. (b) Operation in one neuron. (c) Experimental setup.

layer to the 2nd layer, as shown in Fig. 2.15*a*. The first layer is made of two neurons, each of them getting in 2 inputs and providing 1 output. The two outputs from the 1st layer, are then combined and fed into the second layer which is now made of one single neuron which provides a final output.

Fig. 2.15*b* illustrates one of the neurons with on-chip weighted addition and off-chip optical wavelength converter as a nonlinear function. The all-optical neuron interconnectivity realized with the NL-SOA (Kamelian, NL-SOA) based wavelength converter, who converts the power summation of the weighted multi-wavelength signal into the power at one single wavelength as an optical output and feed to the next layer. These wavelength converters could be integrated on the InP platform in the future.

Fig. 2.15*c* shows a micrograph of the utilized part of the monolithic cross-connect chip and the off-chip optical nonlinear functions with wavelength converter. The blue box indicates where all the three weighting functions for the 3 neurons are placed. The chip is traversed two times without converting the optical signal back to the electrical domain. In the first layer, we use a WDM input In1 and send it to two neurons. After weighting individual channels of the WDM source and combine them again, the output power of the two neurons (O1 and O2) are converted into 2 different wavelengths and combined to create another WDM input

In2 for 2nd layer and send to the neuron 3. The output of this third neuron belonging to the second layer is finally measured with a nonlinear avalanche photodiode (APD) to convert into the electrical domain and analyzed.

2.5.2. Experimental results

The normalized mean root square error (NMRSE) connected to the weighted addition in the photonic integrated chips has already been studied in section 2.1. The transfer function of the nonlinear function is measured by inputting a sine function modified clock signal with 10 GSamples/s. The responses of the wavelength converter are recorded to plot the correlation with the input data. For the optical interconnection from the 1st layer to the 2nd layer, a wavelength converter is exploited, while a high sensitive APD is used at the very end of the output layer so that these two nonlinear functions are observed. Fig. 2.16*a* plots the recorded data respect to the input data, after normalization, from the output of the NL-SOA when it is driven with 80 mA current and signal input power at -10 dBm. The circles are the recorded data, while a solid line is the fitting with hyperbolic tangent function $f=\tanh(\alpha x+b)$, to represent the optical nonlinear function f_{NL}. Fig. 2.16*b* is the transfer function f_{APD} of the APD when input signal with average power at -26 dBm, with fitting the record. These transfer function will be used to calculate the expected output from the final output. The wavelength converter provides a hyperbolic tanh like function with slow slope while the nonlinear function of the APD provides a very rapid change from -1 to 1 as it includes a signal recovery circuit to recover the analog signal back to on-off key signal.

We also study the changes in the nonlinear function by tuning the injection current of the NL-SOA and the power ratio of the CW laser and signal. Fig. 2.17*a*

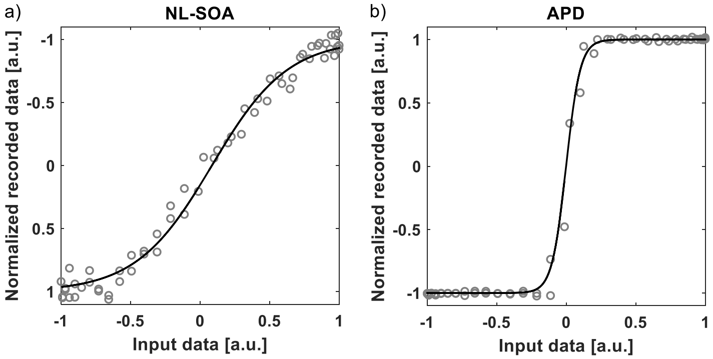


Fig. 2.16 the nonlinear function of (a) the wavelength convertor when injection current is set to 80 mA and (b) the avalanche photodiode when the average power of the input signal is -26 dBm.

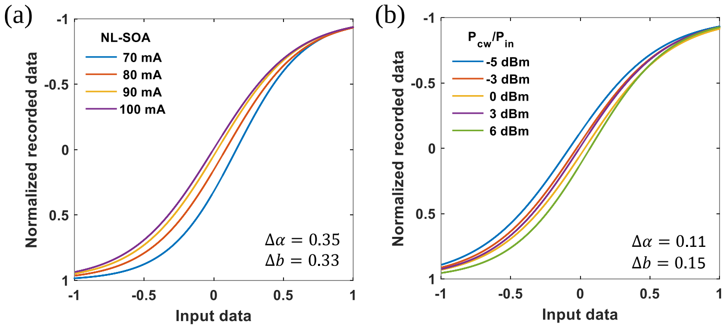


Fig. 2.17 The opitcal nonlinear function changes by tuning (a) the driven current of NL-SOA and (b) the power of the incident signal to the NL-SOA, fitted with function $f_{NL}=\tanh(\alpha x+b)$.

plots the transfer function $f_{\mathrm{NL}}=\tanh(\alpha x+b)$ changes with varying the driven current of the wavelength convertor. The recorded data points are not presented for clarity, and the y-axis is reversed as the output of this wavelength convertor structure is inverted signal. It appears that the slope of the tanh function is decreasing when the current increase, and the nonlinear function shift from the origin as well, resulting in a change of slope of 0.35 and the bias of 0.33. Fig. 2.17*b* presents the variation of the transfer function of the wavelength converter when the power ratio between continue wave (CW) laser and the signal power. This could be done by tuning the laser power or the average signal input power, here by tuning the signal input and fixed the CW laser on -10 dBm at the input of the NL-SOA. The result shows the change of slope α and the bias of function f_{NL} are 0.11 and 0.15, respectively. These results suggest the nonlinear function can be engineered by tuning the driven current the input signal power of the NL-SOA.

After recording operation of the nonlinear function of NL-SOA, we employ this property into the neuron signal processing. In this two-layer all optical neural network, the WDM optical signal is weighted by setting the different gain on each channel, and the summation of the multi-channel signal is converted by the NL-SOA. The outputs from the first layer are fed to the second layer as WDM signal, weighted again with weight SOAs, the final addition is recorded by the APD at the end. In this work, the decorrelated PRBS23 signals x_1, x_2 are modulated on the two input channels λ_1, λ_2, with 10 Gbit/s bitrate, and the weight factors of them in neuron 1 are set to w_{11} and w_{12}, in neuron 2 are set to w_{21} and w_{22}, respectively. The outputs of the first layer are accessed with a pre-amplified linear photodetector to study the performance of the optical NL function. The results after the optical nonlinear functions at the outputs of neuron 1 and neuron 2 in the 1st layer are $y_1 = f_{\mathrm{NL1}}(w_{11}x_1 + w_{12}x_2)$ and $y_2 = f_{\mathrm{NL2}}(w_{21}x_1 + w_{22}x_2)$, respectively, with the $f_{\mathrm{NL}i}$ is the optical nonlinear function of neuron i in the 1st layer. The weight

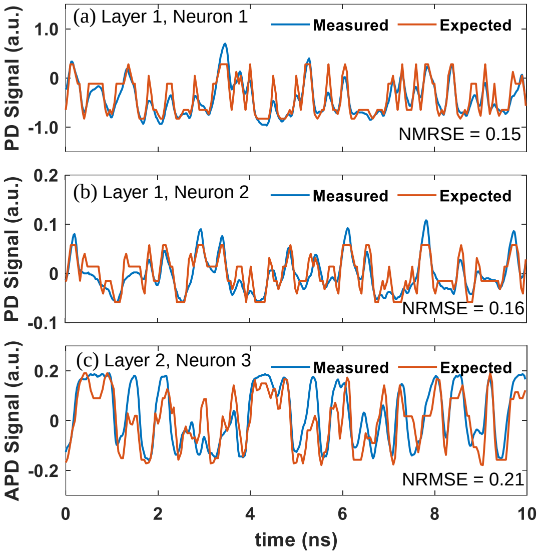


Fig. 2.18 the experimental output signals of the all-optical neural network, at the output of the neurons (a) Neuron 1 in layer 1, (b) Neuron 2 in layer 1, and (c) Neuron 3 in layer 2, with blue line the measured signal and the red line the expectation from calculation.

factors in the neuron at the second layer for λ_3, λ_4 are set to w_{31} and w_{32}. The final output of this neuron 3 at 2nd layer is $y_3 = f_{APD}[w_{31}f_{NL1}(w_{11}x_1+w_{12}x_2) +w_{32}f_{NL2}(w_{21}x_1 +w_{22}x_2)]$, with the nonlinear function f_{APD} of APD at the end.

In section 2.2, we have calibrated the current control to assign the weight factor to the input signal on this chip. The output data with this PRBS signal demonstration is recorded with setting weight factors on chip. Fig. 2.18 depicts the time traces recorded from the output of the neurons from the 1st layer and the 2nd layer. The blue lines plot the measured results from the experimental setup while the red lines plot the expectation values calculated with the original PRBS23 inputs, the assigned weights, and the studied nonlinear functions, as discussed above. Fig. 2.18*a* shows the time trace recorded at the output of neuron 1 in 1st layer, resulting in NRMSE of 0.15, respecting to the expectation values. Fig. 2.18*b* plots the recorded signal at the output of neuron 2 in 1st layer, resulting in NRMSE of 0.16. Fig. 2.18*c* plots the final output signal recorded from neuron 3 in 2nd layer, resulting in NRMSE of 0.21. The experimental results are visibly close to the expectation values. The error increase on this AODNN from the 1st layer to the 2nd layer is comparable to our previous research [69], where an O/E signal conversion is employed after the cross-connect, and this error level would be expected to decrease the performance of a DNN from a computer simulation one about 10% only.

The error from the measurement to the expectation could attribute to the optical signal noise ratio (OSNR) degradation, the mismatch of the nonlinear function responds and the E/O conversion. The OSNR at the input is 42 dB, then it decreases to 35 dB after the chip. The OSNR after NL-SOA is estimated to be 23 dB so that the OSNR after processing by the chip again will be 18 dB. Integrate full layers of AODNN on chip with lower loss is expected to improve the performance. The nonlinear function discussed above is exploited in the expectation value calculation, which might not accurate enough as it is measured with a sine function modified clock signal, which is not the PRBS signal used in the signal processing. This also suggests further optimization can improve the performance of the system. The E/O conversion, i.e. modulation and detection, will also induce error from distortion and detection.

2.6. Conclusion

An InP photonic neural network has been realized and demonstrated based on the SOA technology. The utilized part of the integrated chip provides 4 weighted addition units while the activation function is implemented via software. The on chip SOAs are operated in the linear regime, which reduces the complexity of the weight calibration. This is demonstrated with an NRMSE smaller than 0.08 and a best-case dynamic range of 27 dB.

The Iris flower classification problem has been used as a small-scale example to demonstrate that the photonic neural network concept allows to get as similar accuracy as in electronics. The final prediction accuracy, after a three-layer photonic DNN implemented for an image classification problem, is reduced by 9.2% with respect to the simulated prediction accuracy. A comprehensive error analysis, performed to understand that the on-chip induced impairments, suggest that an on-chip all-optical neural network implementation is expected to improve photonic neural networks performance.

The co-integration of gain elements and filters is foreseen to provide a large dynamic range as well as to enable a route to scalable DNNs. In the future, the photonic neural network structure may be extended to operate as a recurrent neural network, by feeding the optical signal back to the chip. Convolution neural networks are also possible by using the cross-connect scheme and may allow to fully explore the WDM operation (as will be proposed in Chapter 5). However, the real benefit of the photonic integration approach may be demonstrated by running higher bandwidth applications, such as sensing, radio-spectrum manipulation and self-driving cars.

An all-optical deep neural network with integrated InP cross-connect is preliminarily demonstrated with 2 inputs to 2 neurons in the first layer and one neuron at the output layer. The nonlinear activation function with NL-SOA as wavelength convertor is presented to changes the shape of nonlinear function by tuning the driven current and power ratio of CW laser and input signal. The two binary signals are processed with this all-optical interconnect approach, resulting in an NRMSE of 0.21, with is comparable to the E/O conversion approach. Higher accuracy is expected by optimizing the control of nonlinear function. This result opens the way to implement AODNN with photonic integrated circuits on InP platform.

The results suggest that a combination of the weighted addition function with on-chip non-linearities holds the promise to enable further acceleration for computation, which is then demonstrated in Chapter 3.

Appendix 2.A. Simulated Investigation on Crosstalk and Path-loss Impact

In the previous sections, a photonic integrated cross-connect has been used to demonstrate a classification problem, with a crosstalk of -19dB, waveguide losses of 5dB/cm. In this section, the photonic deep neural network with the same structure as in section 2.3 is simulated via VPIphotonics Design Suit, to evaluate the performance of the PDNN as a function of the AWG optical crosstalk, the waveguide losses and the referenced output power.

VPIphotonics is used to simulate the integrated cross-connect-based weighted addition (the synaptic operation) as the basic function of the photonic deep neural network. This software allows for numerical modeling of photonic systems as well as of photonic components within the integrated chips and for different material platforms. The simulated set up is built with symbolic blocks and a hierarchal structure. For the passive elements, we execute the simulation in the

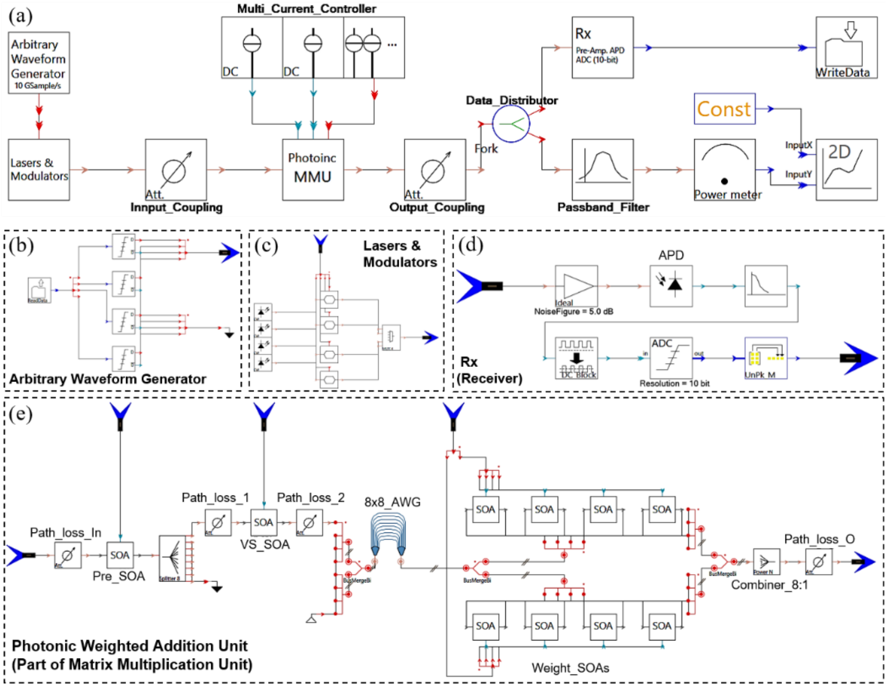


Fig. 2A.1 Photonic deep neural network (PDNN) simulation scheme on software VPIphotonics. (a) System for examining the PDNN. (b) Arbitrary waveform generator. (c) Lasers and modulators. (d) Receiver. (e) One photonic weighted addition unit (part of the matrix multiplication unit, MMU).

frequency domain, while for the active elements, such as the SOAs, the transmission-line model is applied to model them in the time domain [70].

The implemented and simulated setup scheme is showed in Fig. 2A.1. Fig. 2A.1*a* is the complete setup scheme for examining the cross-connect photonic integrated chip shown in Fig. 1c, with similar operating conditions as in the real experiment, for analyzing the integrated SOA-based PDNN. The photonic integrated chip is an 8 inputs × 8 outputs × 8λ cross-connect, but in the experiments, a WDM input is used which contains 4 channels. An arbitrary waveform generator (detailed scheme shown in Fig. 2A.1*b*) is utilized to generate the electrical signal from the data file at 10 GSymbols/s, with 4 DACs with 8-bit precision.

Fig. 2A.1*c* shows 4 lasers and 4 modulators for the optical signal generation of 4 input channels. The WDM input of four channels is generated via these four Mach–Zehnder interferometer-based modulators, with the electrical RF signal coming from the arbitrary waveform generator, and CW lasers at 193.1 THz, 193.5 THz, 193.9 THz, and 194.3 THz. A channel separation of 3.2 nm is used to match the channel separation of the AWG on chip. The input signal is coupled into the photonic matrix multiplication unit (MMU) with a 0 dBm optical input peak power for each channel. The output of the MMU is coupled to the receiver, shown in Fig. 2A.1*d*, which consists of a pre-amplifier with a noise figure of 5.0 dB, an AC-coupled (i.e., with DC-removing block in the simulation) 10 GHz avalanche photodetector (APD), and an analog-digital converter (ADC). The output from the MMU is then coupled to a 0.08 nm optical passband filter to monitor the peak power of one single channel at the output.

The details of the schematic of part of the photonic MMU, i.e., the weighted addition unit, are illustrated in Fig. 2A.1*e* for the weighted addition demonstration. This will be used as the weighted addition part within a three-layer PDNN for demonstrating the iris flower classification. The path loss is the attenuation of the optical signal happening along the waveguide. The input signal is amplified with a pre-SOA and is split into 8 as for 8 neurons. Firstly, we study the performance of one neuron so that only one path carrying one WDM input signal is connected to the next SOA, the input vector selection SOA, that acts as a port selector as shown in Fig. 2.1. The WDM signal is then demultiplexed by an AWG, and the individual channel is weighted by the weight-SOA, and combined at the output of the unit. The parameters used in the simulation for the SOAs are listed in Table 2A.1. The results are reported and explained, as related to the weight calibration and the weighted addition (section 2A.1), and the Iris classification application (section 2A.2) together with the analysis of the impact of the optical crosstalk and the optical path loss.

Table 2A.1. The parameters used in the simulation of SOA.

Parameters	**Value**	**Unit**
Device Section Length	1000×10^{-6}	m
Active Region Type	MQW	
Active Region Width	2.0×10^{-6}	m
Active Region Thickness	250×10^{-6}	m
Active Region Thickness MQW	100×10^{-6}	m
Active Region Thickness SCH	200×10^{-6}	m
Current Injection Efficiency	1	
Nominal Frequency	193.7	THz
Group Index	3.52	
Polarization Model	TE	
Internal Loss	3000	m^{-1}
Confinement Factor	0.3	
Confinement Factor MQW	0.07	
Confinement Factor SCH	0.56	
Gain Shape Model	Flat	
Gain Model	Logarithmic	
Gain Coefficient Linear	4.00×10^{-20}	m^2
Gain Coefficient Logarithmic	6.9×10^{4}	m^{-1}
Nonlinear Gain Coefficient	1.0×10^{-23}	m^3
Nonlinear Gain Time Constant	5.00×10^{-13}	s
Carrier Density Transparency	1.0×10^{24}	m^{-3}
Linear Recombination	1.0×10^{8}	s^{-1}
Bimolecular Recombination	1.0×10^{-16}	m^3/s
Auger Recombination	2.1×10^{-41}	m^6/s
Carrier Capture Time Constant	3.0×10^{-11}	s
Carrier Escape Time Constant	1.0×10^{-10}	s
Initial Carrier Density	8.0×10^{23}	m^{-3}
Chirp Model	Linewidth Factor	
Linewidth Factor	3	
Linewidth Factor MQW	3	
Differential Index	-1.0×10^{-26}	m^3
Differential Index MQW	-1.0×10^{-26}	m^3
Differential Index SCH	-1.5×10^{-26}	m^3
Carrier Density Ref. Index	1.0×10^{24}	m^{-3}
Noise Model	Inversion Parameter	
Inversion Parameter	1.2	

2.A.1. Implementation of weight calibration and weighted addition

For the operation of the SOA-based photonic neural network, a calibration of the weighting is required for correctly assigning the given weight factors to the input data. For this simulation implementation, the weight-SOAs are identical for all the input channels so that we demonstrate the weight calibration on one of the input channels. For the weight calibration, the input can be a non-return-to-zero on-off keying (NRZ OOK) signal or multi-level data input. As the weighting of the input data is performed after the AWG, the fixed optical crosstalk from the AWG will influence the output optical signal. We consider two extreme conditions for the optical crosstalk level: when switching ON (injection current at 70 mA) all the weight-SOAs, the optical crosstalk coming from the adjacent channels is expected to be maximum ($XTalk_{max}$), while when all the weight-SOAs are OFF (zero injected current), the induced optical crosstalk by that the corresponding channels will be the minimum ($XTalk_{min}$). Due to the complexity of operation conditions, we consider the average between these two scenarios in order to generate the weight control curve, in order to minimize the error induced by the optical interference.

The crosstalk in the AWG in the photonic MMU (see Fig. 2A.1e) is set at −20 dB, as experimentally measured in [59]. Firstly, when all the weight-SOAs are set to OFF ($XTalk_{min}$), but one of these weight-SOAs is injected with currents from 0 mA to 70 mA, we record the signal peak power at channel 1, 193.1 THz, from the monitoring power meter as shown in Fig. 2A.1*a*. Then, we also record the signal power when all the weight-SOAs are set to ON ($XTalk_{max}$), and one of these weight-SOAs is injected with currents from 0 mA to 70 mA. The blue and red solid lines in Fig. 2A.2*a* plot the simulated result in the condition of $XTalk_{min}$ and $XTalk_{max}$, respectively. For comparison, we also superimpose the measured curves in both cases: the blue crosses curve represents the experiment points with all SOAs ON, and the red triangles curve plots the experimental results with all SOAs OFF.

The curve trends, in the case of simulation and experimental results, are very similar. We then scan the optical crosstalk level to investigate the influence of the optical crosstalk on the peak power curves. In Fig. 2A.2*b*, the blue, red, yellow and violet solid lines show the peak power on channel 1 (averaged from the curves shown in Fig. 2A.2*a*), where the simulated crosstalk for the AWG is set to −15 dB, −20 dB, −25 dB and −30 dB, respectively. It is visible that higher crosstalk will induce greater oscillation when tuning the injection current at the weight-SOA. The oscillation might be due to the interference between the crosstalk and the signal in the desired path. The experimental result is also presented with red

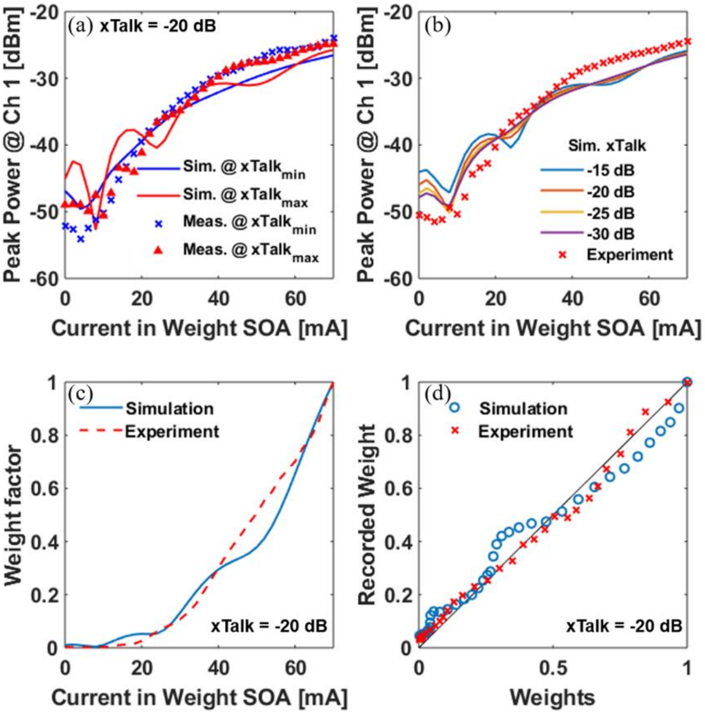


Figure 2A.2 Weight calibration. (a) Peak power of channel at 193.1 THz with minimum crosstalk (blue) and maximum crosstalk (red), in simulation (solid line), or experiment (cross/triangle points), versus the injection current at the weight-SOA. (b) Mean peak power of channel at 193.1 THz from the two curves obtained in (a), and with simulated crosstalk at −15 dB, −20 dB, −25 dB and −30 dB. (c) Weight control curves, in simulation (solid line) and experiment (dash line), with crosstalk of −20 dB and reference power level at −25 dBm. (d) Correlation between the weight assigned by the weight-SOA and the obtained weight at the output, in simulation (blue circles) and experiments (red crosses). The black line is a reference line for perfect matching.

crosses, which is the mean value of the experimental results shown in Fig. 2A.2*a*. The plots indicate that a dynamic range wider than 25 dB is possible. The slight difference between simulations and experiments may be attributed to the difference in gain efficiency as hypothesized for the SOA modeling. The weight control curve in Fig. 2A.2*c* is generated by the power control curves in Fig. 2A.2*b*, with reference weight ‘1’ level at −25 dBm optical input power, which is the signal peak power when injection current of the weight-SOA is set at 70 mA. The weight calibration curves show two semi-linear operation regimes, both for the simulation (Fig. 2A.2*c*, blue solid line) and the experiment (Fig. 2A.2*c*, red dashed line). These two regions correspond to the two different SOA operation regimes: the transparency operation and linear amplification. After the weight calibration,

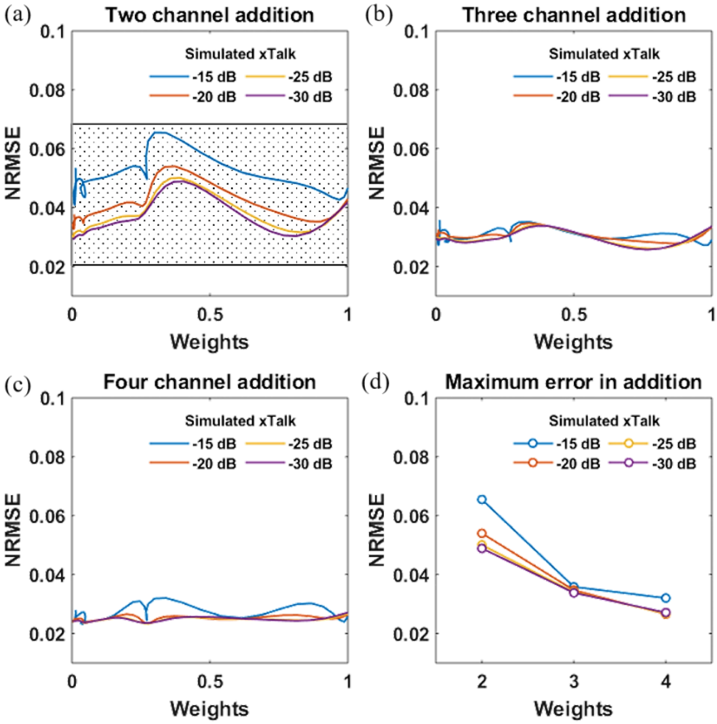


Figure 2A.3 (a) Two-channel weighted addition, (b) three-channel addition, and (c) four-channel addition when tuning weights in channel 1 and fixed weight on other channels; (d) Maximum error versus the number of channels in weighted addition. Optical crosstalk levels are −15 dB (blue), −20 dB (red), −25 dB (yellow) and −30 dB (violet).

we obtain the correlation between the assigned weight and the obtained one for the simulated and the experimental operation in Fig. 2A.2*d*. An error lower than 0.12 for the simulation results is obtained, when compared to the reference perfect linear relation as shown by the black line.

The weighted addition corresponds to the linear operation part in a neuron. The performance of the weighted addition is of importance for the signal processing in a neural network. To estimate the impairments induced by the weighted addition, we calculate the normalized root mean square error (NRMSE), i.e., the discrepancy between the measured data and the expected data. We use the calibrated weight control curve to set the weight factors for different input channels, and calculate the NRMSE while tuning the weight factor from 0 to 1. Fig. 2A.3*a* plots the results of two-channel weighted addition, where channel 2 is fixed to the weight ’1′, while the weight for channel 1 is tuned over the overall range from 0 to 1. We also change the optical crosstalk in the chip to see the impact of the optical crosstalk on the weighted addition. The blue, red, yellow and violet lines show the error changes when the crosstalk is set at −15 dB, −20 dB, −25 dB and −30 dB, respectively. The shaded area shows the error range obtained from the experiments for two-channel addition. The error variation is attributed to the calibration of the weight control, as already anticipated in the weight factor curve in Fig. 2A.2*c*. The error related to the weighted addition operation increases when the induced optical crosstalk is greater, as a high level of crosstalk eventually results in a lower dynamic range. The same high crosstalk level enhances the peak power oscillation recorded for generating the weight control curve, resulting in severe error variations, as already anticipated in Fig. 2A.2*d*. Nevertheless, this matches perfectly the error variation window we found for our experimental results (see the dashed box in Fig. 2A.3*a*). The same analysis is done while changing the number of channels added to the WDM input. Fig. 2A.3 *b,c* plot the resulting errors for three- and four-channel weighted additions, respectively. A visibly smaller error is presented for three-channel and four-channel weighted additions. The effect of the optical crosstalk is reduced as the oscillation power caused by the optical crosstalk is relatively smaller with respect to the dominating signal power coming from the addition of all the input multiple signals. This suggests that the higher the number of inputs into the neuron, the better the accuracy when operating within the available power budget. Finally, Fig. 2A.3*d* summarizes the obtained results, by plotting the maximum errors versus the number of channels in weighted addition.

2.A.2. Image classification via a three-layer PDNN

To investigate the performance of a complete neural network based on the combination of the AWG and SOA technology, for a broadcast-and-weight architecture, we implement and simulate an image classification problem, namely the iris flower classification problem, which has been reported to be able to be solved by using a deep neural network (DNN). The iris database includes three classes (Setosa, Versicolor, and Virginica) of 50 instances each [66]. Per each instance, the iris flower category is identified by observing four of its attributes: length and width of its sepals and petals. For this demonstration, we have executed the training of this DNN via the simulation platform *Tensorflow* [67], where we have used 120 instances as a training database. In order to make use of 4 weighted addition circuitries already available on chip and per layer, a feed-forward network made of 2 hidden layers with 4 neurons each and an output layer with 3 neurons (see Fig. 2A.5*a*), is trained on a computer. The attributes are encoded into 2^6 optical

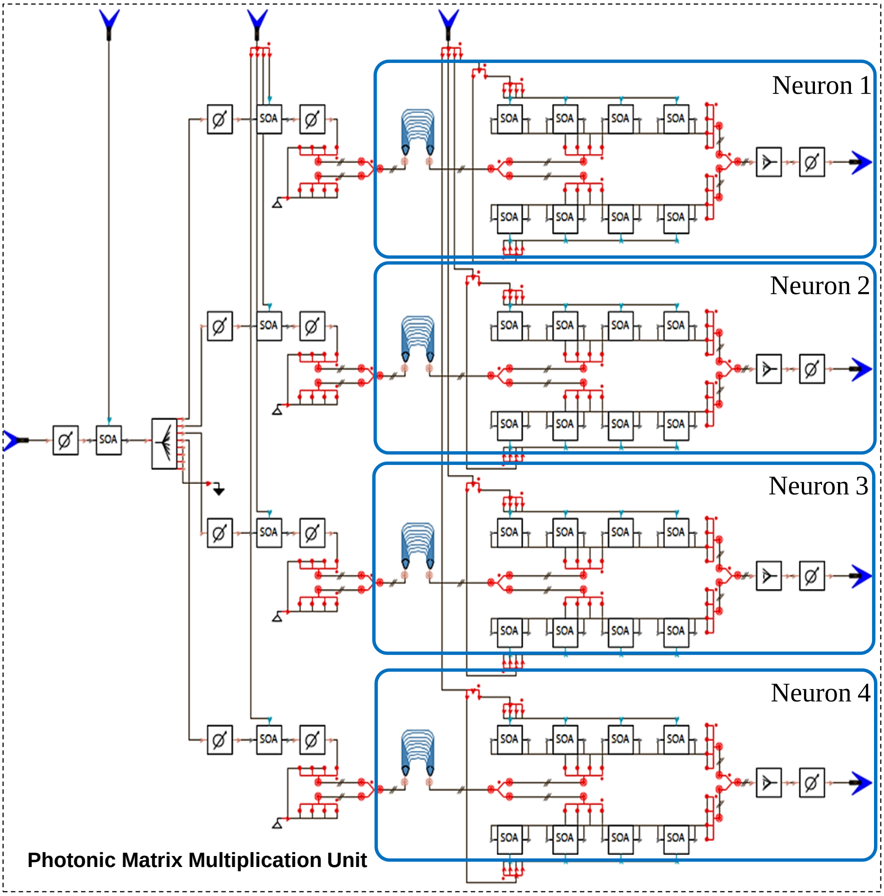


Figure 2A.4 The simulation structure of one layer of neurons.

power levels at the photonic MMU input. The trained weight matrix is mapped to the matrix multiplication on the photonic components. The simulated structure of one layer of neurons is shown in Fig. 2A.4, which is used to replace the photonic matrix multiplication unit in Fig. 2A.1*e*. The same chip is indeed capable of eight channel inputs, but we used four inputs for this classification problem. A total of 16 weight-SOAs in this matrix multiplication unit are used to assign the trained weight matrix from the trained DNN model to the PDNN. The hyperbolic tangent activation function is implemented offline after the O/E conversion. The output from the first hidden layer serves as input to the second layer (via the arbitrary waveform generator) and the output from the second hidden layer serves as input to the third (output) layer. Finally, the output of the third layer, after the *SoftMax* transfer function, $P(y = j) = e^{y_j} / \sum_{j=1}^{n} e^{y_j}$, provides the predicted probability of the output samples y of belonging to class j.

Fig. 2A.5*b* presents the output data at the output (i) of the 1st neuron in the 1st layer, (ii) of the 2nd layer and (iii) of the 3rd layer, with the blue line being the simulation results and red line the expectations and resulting in errors of 0.123, 0.051 and 0.055 respectively. These errors represent the performance of layers of photonic neurons. Higher error at the first layer may be due to the high optical signal noise ratio (OSNR) required multilevel encoding of the input signal, while a better performance at the 2nd and 3rd layer is attributed to the filtering of the signal level into lower levels after the first hidden layer. Also, the output of the first layer appears to be the most important for this classification problem, and therefore the utilization of three layers is slightly overstated.

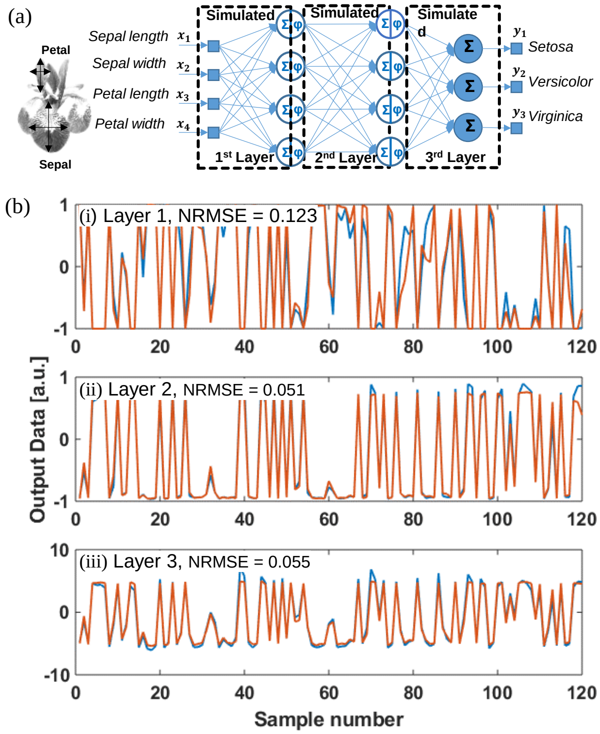


Figure 2A.5 (a) Trained 3-layer deep neural network (DNN) employed to solve the iris flower classification. (b) Output data obtained from Neuron 1 at (i) Layer 1, (ii) Layer 2, and (iii) Layer 3, with calculated errors between simulated computation (blue line) and the expected computation (red line).

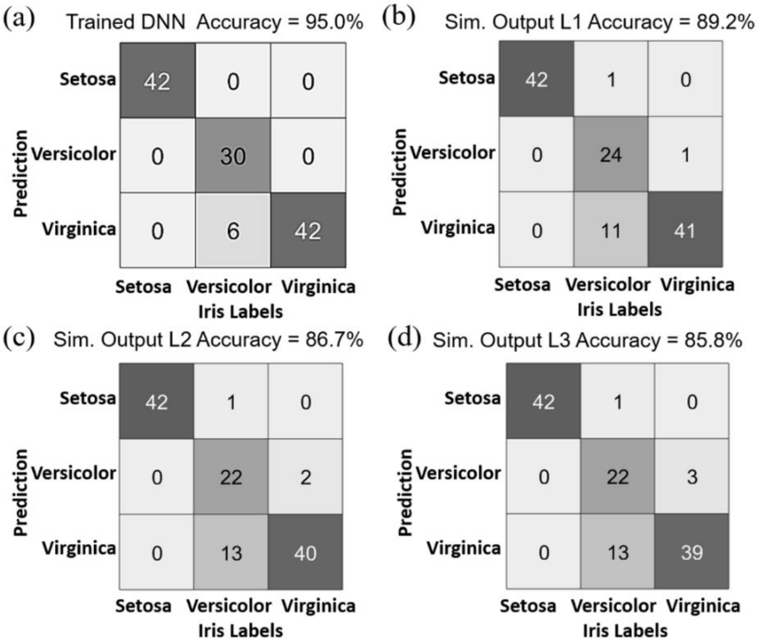


Figure 2A.6 (a) Label prediction of the trained DNN, indicating an accuracy of 95%. (b) Simulated image prediction using photonic DNN as the 1st hidden layer, with an accuracy of 89.2%. (c) Simulated label prediction using photonic DNN as the 1st and 2nd hidden layers, with an accuracy of 86.7%. (d) Simulated label prediction of the 3-layer photonic DNN, with an accuracy of 85.8%.

The correlation matrix between the prediction and the labels of the samples is used to show the final accuracy obtained via the multilayer photonic neural network (see Fig. 2A.5*a*). We consider three cases for the sake of understanding the influence of the photonic layer implementation. In Fig. 2A.6*a* we display the prediction accuracy as coming from the trained DNN on a PC. This is calculated to be 95% since 6 out of 120 iris flower instances are falsely predicted.

The DNN is simulated after adding, time by time, the photonic deep neuron network layers. The prediction accuracy decreases as the number of layers of the photonic neural network increases. The accuracy changes from 89.7% when the 1st layer is substituted with a photonic layer (Fig. 2A.6*b*), down to 86.7% when both the 1st and the 2nd layers are substituted with photonic layers (Fig. 2A.6*c*), and down to 85.8% when all 3 layers of matrix multiplications are computed via three photonic layers (Fig. 2A.6*d*). This may be due to the error accumulation which causes prediction accuracy degradation. Furthermore, the simulation result aligns well with the experimental result trend as shown in Fig. 2A.7*a*.

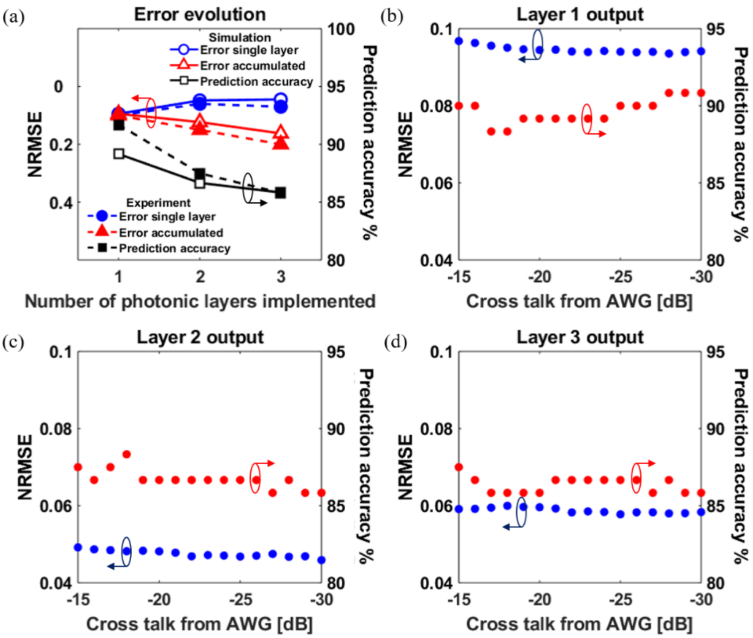


Fig. 2A.7 Error evolution: (a) Normalized root mean square error (NRMSE) versus the number of implemented photonic layers in simulation (solid line filled points) and experiment (dashed line open points. Crosstalk tuning: The induced error (blue circles) and the final prediction accuracy (red circles) versus the crosstalk from AWGs, recorded simulation results from (b) output of layer 1, (c) output of layer 2, and (d) output of layer 3.

Fig. 2A.7*a* plots the error evolution with an increasing number of the photonic layers. The solid lines with open symbols represent the results from the simulation and the dashed lines with filled symbols represent the experimental results. The circles show the error induced by each single layer on the 3-layer network, where the errors keep staying almost at the same level, about 0.07 in the simulation and 0.08 in the experiment. The triangles plot the accumulated error from layer to layer, which increases from 0.1 to 0.18 for simulation and from 0.10 to 0.20 in the experiment. The squares represent the final prediction accuracy as we calculated from the correlation matrix, which decreases from 89.2% to 85.8% as shown in Fig. 2A.6 for simulation and from 91.2% to 85.8% for experiments, as shown in section 2.3. The experimental results show great agreement with the simulations, which means that investigating the performance while changing some of the parameters involved in the photonic integrated circuit will help to get some insight into the photonic chip architecture and scalability. From the perspective of the final prediction accuracy and error induced by the photonic neural network chip, the impact of the optical crosstalk from the AWG and the waveguide crossings are investigated. The crosstalk of the AWG is tuned from −15 dB to −30 dB with 1 dB steps and implement the 3-layer neural network after generating the weight calibration curves as reported in section 2A.4. Fig. 2A.7b plots the results at the output of layer 1, with the blue line representing the average NRMSE from 4 neurons at layer 1 and the red line plotting the variation of the final prediction. Similarly, Fig. 2A.7 *c,d* illustrate the average NRMSE and the final prediction versus the optical crosstalk in the layers. The error induced by the chip is almost in the same range for different crosstalk values, though it slightly reduces when the crosstalk decreases. The prediction accuracy for Layer 1 in Fig. 2A.7*b* shows a stronger crosstalk dependency; a smaller optical crosstalk at Layer 1 provides a better prediction accuracy. This may be related to the fact that the first layer operates on high resolution multilevel input signals, which require a higher optical to signal ratio available. A better accuracy also appears when the crosstalk is high, i.e., near −15 dB. This might be attributed to the errors leading the prediction of the flower label to a different minima location, i.e., to changes of the state of the network, as will also be found in next section. Fig. 2A.7*c* shows a flattened accuracy for optical crosstalk smaller than −20 dB. Fig. 2A.7*d* shows an even more flattened accuracy level as the variation of the induced error is smaller. The accuracy level is maintained from the 2nd layer onwards.

2.A.3. Energy consumption versus physical layer impairments

The performance of the PDNN is expected to be influenced also by the reference energy level used to operate the optical engine and by the loss on chip. Therefore, we study the performances of the PDNN by executing the iris classification problem, while tuning the reference power level of weight factor ‘1’ used in the integrated circuit, i.e., while tuning the current used in the SOAs, as well as while scanning waveguide losses for the optical paths. A more complete analysis would require the inclusion of all patch passive losses, however we consider these simulations as an initial step to further understanding the Photonic deep neural network operation capabilities. This analysis is carried out to understand opportunities for energy savings and best chip physical layer characteristics, which still guarantee a high level of prediction accuracy. In particular, for this analysis, we consider only the waveguide loss as the main loss component as this is true for large size PIC. Therefore, the NRMSE is calculated at the output of each photonic neuron layer, as well as the prediction accuracy obtained when involving this layer in the 3-layer DNN for the iris flower classification, and we provide 3D color maps of error and accuracy as a function of the scanning waveguide losses and energy consumption on different reference power levels.

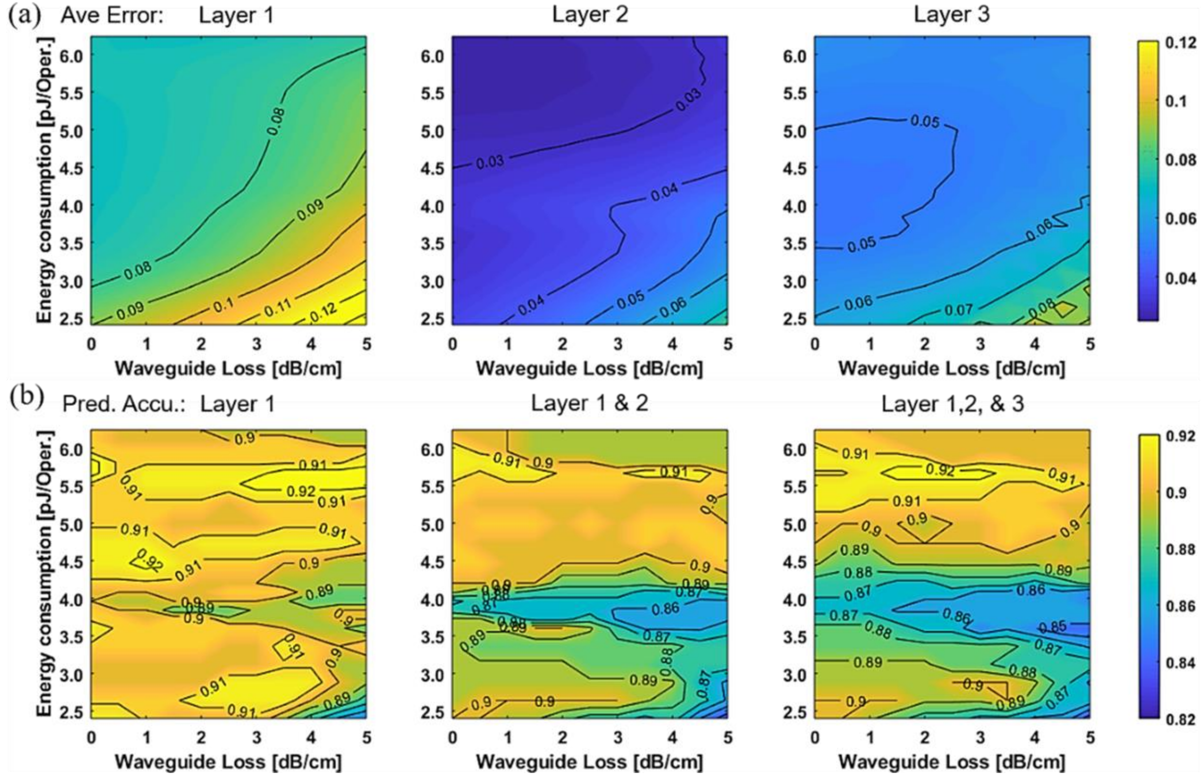


Figure 2A.8 Investigation performance of the PDNN on computing energy and waveguide loss. (a) Calculated average NRMSE from output data obtained from Layers 1–3; (b) Corresponding prediction accuracy when Layer 1, Layers 1,2, and Layers 1–3 are implemented with photonic neuron layers.

Fig. 2A.8*a* illustrates the average errors obtained at the output of the 1st layer, the 2nd layer and the 3rd layer as a function of the waveguide loss and the energy consumption. It can be observed that fewer losses allow less energy consumption, for the same error level. This suggests that by only improving the waveguide loss on chip we can double the energy savings. The induced error from the photonic DNN is expected to be greater when the waveguide losses are higher and the energy consumption per operation at the matrix multiplication unit is lower, as the dynamic range is not enough to be able to distinguish multilevel data. On the contrary, smaller error values are observed with lower waveguide loss and higher energy used for the weighted addition operations (see moving from lighter color to darker color, from bottom right to top left side in Fig. 2A.8*a* and for each layer). For Layer 1 this is more evident and is coherent with previous conclusions. It is not surprising that we need to either tune the reference power to higher levels or reduce the waveguide loss to obtain smaller errors at the neuron signal processing. Furthermore, the final prediction accuracies for the cases when Layer 1, Layers 1 and 2, and all three layers are implemented by using the photonic integrated chip are shown in Fig. 2A.8*b*. The yellow color corresponds to a higher prediction accuracy, while the blue color corresponds to a lower prediction accuracy. The prediction accuracy results do not show the same trend as shown in Fig. 2A.8a, i.e., the trend for the error induced on the layer operations, which indicates that an induced error is not necessarily reducing the performance of the photonic neural network. The result for Layer 1 shows that a good prediction is obtained for an error smaller than 0.09 and in that region the accuracy remains generally very stable, while the accuracy for higher error levels is variable, and generally worse. This suggests that there is a certain maximum level of error we should never cross at the first layer for always guaranteeing a good accuracy. The prediction mapping from the implementation of Layers 1 and 2 shows a slight decrease in the accuracy as the error accumulated from the previous layer. In the case of adding the contribution of Layer 3, it is the result of the small error accumulation as well, for the final prediction accuracy on this 3-layer photonic neural network system.

However, it is evident that the two different regions are delineated when more error is accumulated from layer to layer. In particular, for the complete 3-layer photonic neural network, the best performance condition (accuracy = 92%) is found when the energy efficiency is around 5.6 pJ/operation and the waveguide loss ranges between 1.5 and 3.5 dB/cm. However it is possible to distinguish two areas where the accuracy is already higher than 89%: (1) the total energy consumption is above 4.5 pJ/operation, irrespective of the path loss; (2) the area at the left down corner where the averaged energy consumption is around 2.8 pJ per operation and the loss covers almost the full considered range (up to 4 dB/cm).

The region (1) performs well due to a higher signal power with higher power consumption on the system with smaller errors induced and accumulated. We believe the region (2) appears due to the presence of more local minima, whose presence is determined by the combination of path losses, power level and optical crosstalk.

Furthermore, the level of noise present in the network makes it a stochastic network where the intrinsic noise is supposed to provide better accuracy. Noise might play a positive role for low power levels as a good prediction is presented. However, this behavior has to be further explored for quantification. The identification of small error regions and their slight influence on the final prediction accuracy, as well as the maximum level of error at the first layer shown in Fig. 2A.8 suggests that the PDNN might be further scaled up, with prior physical parameters and error optimization.

2.A.4. Conclusion

The integrated photonic neural network is also simulated as a weighted addition, combined with an offline hyperbolic tangent nonlinear function on the simulation platform *VPIphotonics* as shown in Appendix 2A. The weight calibration and weighted addition with different crosstalk of photonic integrated AWGs and SOA-based cross-connects is investigated. The error from the weighted addition is found to decrease when the numbers of input channels increase, so that a high number of input channels is beneficial for the implementation of the PDNN (this is also further analyzed and validated in the next chapter). A trained 3-layer DNN is implemented by reconfiguring the weight setting on the subnetwork and feeding the layer output to the next layer. In particular, the performance is simulated with different values of crosstalk, energy consumption per operation, and waveguide loss. The experimental results are in agreement with the simulation results, meaning that the implemented simulation offers a solid base for further study of scalability for this kind of network architecture. The results show that the photonic DNN is robust to the noise added during the signal processing. The error induced by the first layer is greater than the next ones, due to the higher resolution multilevel encoding at the input layer with respect to the resolution at the 2nd and 3rd layer, but the error is not necessarily degrading the performance for a maximum allowed error. The performance analysis as a function of the path losses suggests the photonic neural network could be further optimized for lower power consumption. These results provide initial insights for the design of scalable photonic neural networks to a higher dimension for solving higher complexity problems.

Chapter 3

SOA-based Photonic Integrated All-Optical Deep Neural Network

In this chapter, the structure of an all-optical neural network with SOAs in linear and nonlinear operation region is proposed. The co-integration of SOA-based synaptic operations, in the form of a combination of SOAs and AWGs, together with a nonlinearity, in the form of a wavelength converter based on cross-gain modulation, are presented. The envision of the all-optical SOA-based neural network structure is introduced in section 3.1. And the integrated version of AON is experimentally demonstrated in section 3.2. The nonlinear function and the operation of the fully monolithically integrated all-optical neuron are analysed. Then in section 3.3, the all-optical network is exploited and simulated to solve the handwriting digit classification MNIST problem to evaluate final accuracy. Finally, in section 3.4, the energy consumption of the complete end-to-end system is analysed.

3.1. All-Optical Deep Neural Network

The overall envisaged all-optical deep neural network scheme based on the use of the cross-connect circuitry is depicted in Fig. 3.1. The wavelength division multiplexed (WDM) signal from N input neurons (one wavelength from each input neuron) is fan-out towards the following M neurons of the first hidden layer. At each i-th neuron of this layer (highlighted through an orange box), the multiple-wavelength signal is demultiplexed into N signals, which are multiplied, via an SOA, by the weight $w_{i,j,k}$ of the i-th neuron, from the j-th axon (λ_j) and in the k-th layer. The weighted signals, being encoded in different colors, are then summed up via an AWG-based multiplexer. This first circuitry (black dashed box in Fig. 3.1) corresponds to the linear part of the i-th neuron, whose output is sent out to a nonlinear function block $NL_{i,k}$ of the same neuron of the k-th hidden layer.

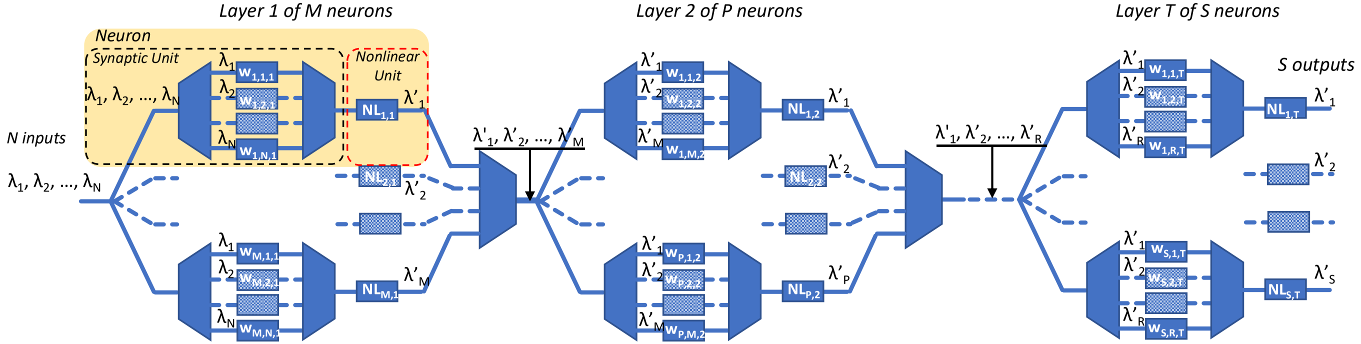


Fig. 3.1 Schematic of the foreseen deep neural network based on SOA-based weight and non-linear functions. NL: non-linear function. $w_{i,j,k}$ = weight of the *i*-th neuron, of the *j*-th axon (λ_j) and of the *k*-th layer.

This is realized via an SOA (red dashed box in Fig. 3.1), where the enabled cross-gain modulation (XGM) is used to output a wavelength-converted light, modulated by the total power of all WDM channels at the input of the SOA-based wavelength converter (SOA-WC), for the conversion into a different single wavelength, λ_i', which represents the output of this neuron. The outputs from all the neurons of the *k*-th hidden layer are then combined and broadcast again towards all the neurons of the next hidden layer, and so on and so forth. It is important to note that the shuffle network here is obtained by combining AWGs and one big 1:*M* splitter (for example moving from the first to the second hidden layer), in the place of *M* times AWGs, which would deteriorate crosstalk, as well as introduce a deleterious path dependent loss.

Here the SOA technology is exploited in combination with the array waveguide grating (AWG) technology for multiple reasons: The optical amplifiers are employed for setting the weight matrix and providing on-chip gain for scalability, while the AWGs are used to filter out the out-of-band noise built up by cascading multiple stages of SOAs, in order to increase the weight resolution, as well as to carry out the needed multiplexing and demultiplexing functions. In Chapter 2, the synaptic operation using an 8×8 InP SOA-based cross-connect chip is demonstrated, followed by an array of photodetectors to further process the signals in the electrical domain and to perform an analysis of the sources of error. Indeed, a reduction in accuracy happened, which was dominated by the electro-optical conversions needed to move from the optical linear function to the electrical nonlinear function, as well to progress from one layer on chip to the next one, which suggested to move to an all-optical approach, shown in section 2.5.

In this section, this investigation is extended to other sources of errors versus the scalability properties of the linear neuron, specifically the crosstalk as a function of the number of the channel inputs (axons). The optical crosstalk, coming from the AWGs, limits the linear circuitry scalability as soon as the number of neuron inputs (channels) increases. For this reason, the normalized root mean square error (NRMSE) of the synaptic unit of the neuron is calculated, as shown in Fig. 3.1 (black dashed box), while tuning the total input power and the total number of neuron inputs, for a channel spacing of 100 GHz. This has been analyzed via the VPIphotonic Design Suit (using parameters as detailed in Appendix 2.A). In Fig. 3.2*a* the colored lines represent the error obtained after the synaptic unit for 4, 8, 16, 32 and 64 inputs/neuron while tuning the total input power of the WDM input to the neuron from -25 dBm to 20 dBm. The results show that there is an optimized optical power operation point for reaching the minimum error. It is notable how this point shifts down-right side for a larger number of neuron inputs. To better visualize and explain this trend, the same NRMSE is

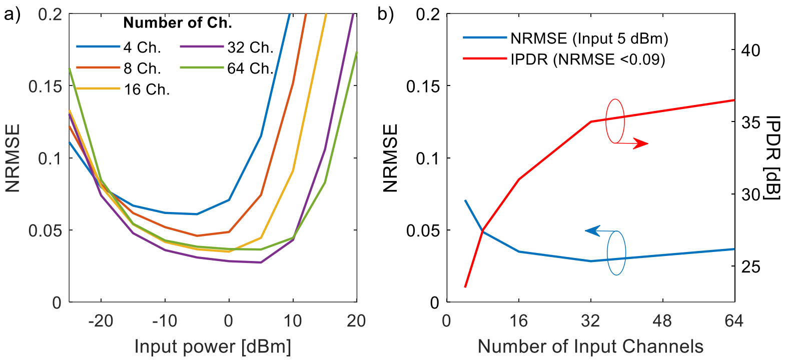


Fig. 3.2 (a) Error obtained when tuning the total input power to the neuron from -25 dBm to 20 dBm per different input channel numbers. (b) Error variation (5 dBm input power, blue line) and IPDR (NMRSE < 0.09, red line) when changing the numbers of input channels from 4 to 64 channels. The AWG channel spacing is set at 100GHz.

plotted as a function of the number of the input channels for a fixed input power of 5 dBm at the AWG input (see Fig. 3.2*b*). When increasing the number of neuron inputs, the error decreases, as shown in blue line, while it starts slightly increasing only for a number of channels higher than 32: The vertical scalability of the neural network (height), and therefore a higher channel number, results in an increase of the resolution of the linear summation output, since more channels at the input contribute to increasing the total number of the output signal levels within the same dynamic range, resulting in a smoother output signal pattern. In particular, the error is found to increase for 64 channel inputs due to the limited modeled SOA bandwidth (71.5 nm 3-dB gain bandwidth), in fact 64 channels spaced 100 GHz already fill up 51.2 nm bandwidth. The red line in Fig. 3.2*b* plots the input power dynamic range (IPDR) as a function of the number of the input channels per neuron and for an NMRSE < 0.09: for this level of error, previously in Chapter 2, a 3-layer neural network is shown resulting in <5% degradation of the prediction accuracy for an image classification problem. The IPDR increases from 25 dB to 36 dB, which is partly attributable to the large SOA linear regime (-5 dBm input saturation power) but also to the fact that, with the increasing number of input channels, the power fed to the individual weight SOA will be much lower than the input saturation power, making the SOAs working in the linear regime for a wider input power range. The trend slows down when the number of channels approaches 64 since it come closer to the bandwidth edges of the SOA.

3.2. SOA-based Integrated All-optical Neuron

In this section, the possibility of realizing an all-optical SOA-based neuron is investigated, to realize multi-layer neural networks and avoid electro-optical conversions for improving energy efficiency, while guaranteeing still a good accuracy. To this aim, the optical output of the synaptic operation is input straight to an SOA-based nonlinear function. The exploitation of SOA-based circuits for both the linear and nonlinear functions of an artificial photonic neuron enable the monolithic integration of both functionalities to overcome optical loss issues deriving from a hybrid approach, as shown in section 2.5. After the explanation of the experimental set up (section 3.2.1), the nonlinear function based on a wavelength converter (section 3.2.2), and then the overall performance of the complete neuron is investigated, integrating the optical linear neuron with the SOA-WC optical nonlinear function (section 3.2.3).

Before describing the experimental measurements and simulations in the following sub-sections, it is important to discuss the assumptions made on any four-wave mixing effect happening within the nonlinear SOA. Depending on the wavelength separation, input power and the number of WDM channels, the FWM inside the SOA may have a non-negligible influence on the overall performance. However, this is not considered in the simulation, neither is observed in the experimental phase. In fact, this effect is neglectable when the detuning between the probe channel and the pump frequency $\Delta f >> 1/(2\pi\cdot\tau)$, where τ is the carrier lifetime of the used SOA. In this work, the carrier lifetime is estimated to be 200 ps worse case [71], and the channel spacing at the input is 100 GHz and 400 GHz, for the simulations and the experimental work, respectively, which results in detuning that is far greater than $1/(2\pi\cdot\tau) \approx 1$ GHz. By exploiting the methods in Ref. [72] and [73], one can estimate that the conjugate signal generated by the FWM effect has a power of the order of < -64 dBm when the detuning used in simulation is >100 GHz, which is even lower than the spontaneous emission noise at the neuron output. Moreover, in order to suppress the FWM for larger number of input channels, the total input power of the neuron is controlled by defining an appropriate scaling of the neural network. In our approach, the network size is considered scaling up with input channels N and with the same number of neurons M. In this way, the total input power to the neuron will stay constant when scaling N and M, i.e., for each channel power p_0, the total input power at the layer input $N\cdot p_0$ will be split towards M neurons as $N\cdot p_0/M$. When setting $N=M$, the total power to each neuron (yellow box in Fig. 3.1) will be p_0, and the power for each channel p_0/N. For such a condition, the FWM effect results reduced due to the decrease of the input signal power, as well as to further detuning of the individual input channels. Finally, using unequal channel spacing at the WDM inputs [74],[75], the FWM effect can be further eliminated.

3.2.1. Experimental setup

The experimental setup used for the SOA-based all-optical neuron investigation is depicted in Fig. 3.3*a* with a micrograph of the fabricated chip shown in Fig. 3.3*b*. A four-channel WDM optical input is composed of signal wavelengths set at 1544.0 nm, 1546.0 nm, 1551.0 nm and 1554.0 nm in order to match the nominal 3.2 nm channel separation of the on-chip AWGs with a 3-dB bandwidth of 0.8 nm and to maximize each channel optical power output. The input is modulated with PRBS on-off keyed (OOK) data, generated by the arbitrary waveform generator (Tektronix, AWG7122B), and sent to the input of the integrated all-optical neuron after de-correlation. The WDM optical inputs are equalized and set at -12 dBm power per channel.

Inside the neuron in Fig. 3.3*b*, the inputs are amplified with a booster SOA, which is utilized to optimize the total power at the input of the weighting SOAs.

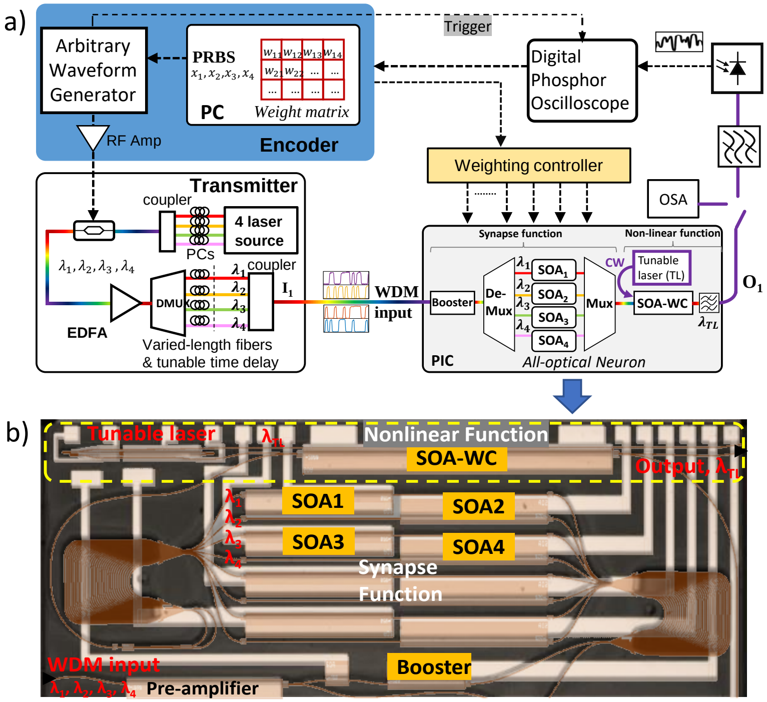


Fig. 3.3 (a) Experimental setup for SOA-based all-optical neuron investigation. (b) Micrograph of the integrated all-optical neuron.

Then the signals are weighted with individual SOAs after channel demultiplexing by the AWG and combined again with an AWG-based multiplexer and fed to the SOA-WC based optical nonlinear function, whose pump laser is fully integrated on chip. This provides a converted output at 1549.0 nm, chosen to be close to the center of the WDM channel bandwidth for optimizing the wavelength conversion.

The weighting SOAs are controlled by a weighting current controller (Thorlabs, MLC8200CG), with 50 µA resolution, to provide the weights in 10-bit precision, which exceeds the required precision for image classification. The current synapse control is envisioned to be realized by means of an FPGA controlling multichannel current drivers [59], when further scaling the number of neuron synapses. In the future, the parallel development of ultra-compact driver ICs, of new electronic interface techniques and of cleaver electrical control schemes seems to be a viable route towards enabling control of larger size photonic networks on chip.

The individual weight SOAs are calibrated to compensate for the wavelength conversion non-uniformity among the different channels: This calibration happens prior to the assignments of the actual weighting factors. The noise figure and the saturation output power of these SOAs are 7 dB and 8 dBm, respectively. The output of the all-optical neuron is detected by a linear avalanche photodetector (PD) and the time trace is recorded by a digital phosphor oscilloscope. The performance of the neuron is again evaluated by calculating the NRMSE between the recorded and the expected time traces at the output of the NL function, calculated using the reference pre-recorded inputs. The synaptic operation of the neuron can be expressed as a weighted addition of parallel inputs: $y=\sum w_i x_i$, where w_i is the i-th weight element for input x_i, and the final output of the neuron is: $o=\varphi(y)$, where φ is the nonlinear transfer function of the SOA-WC.

3.2.2. Integrated SOA-based non-linear function

The wavelength converter is the nonlinear device that exploited as an optical nonlinear function within the neuron. The SOA-WC is also integrated on the InP platform, with an on-chip tunable laser [76]. The integration of the all-optical nonlinear function allows us to demonstrate a monolithically integrated SOA-based all-optical neuron [77], as shown in Appendix 3.A. In order to measure only the transfer function of the nonlinear part working at first as a simple inverter, the PRBS OOK input of the neuron and the output of the SOA-WC is recoded. The correlation map of the two is the nonlinear transfer function (NL-TF), which can be used to calculate the expected output for the entire all-optical neuron. The blue line time trace in Fig. 3.4*a* plots the pre-recorded 2 Gbit/s single channel input signal. Fig. 3.4*b* presents the detected output of the SOA-WC based NL-TF

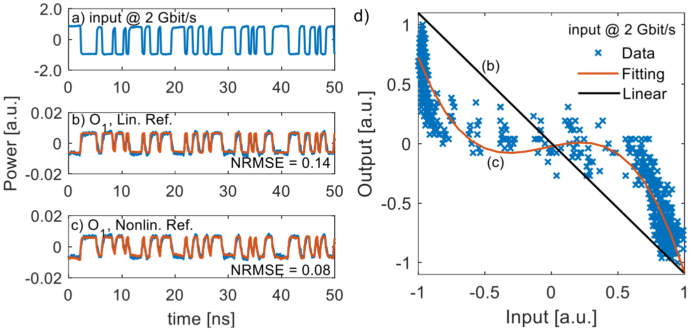


Fig. 3.4 (a) the input data in channel 1 at 2 Gbit/s. (b) The recorded output (blue) comparing to the expectation (red) with linear transformation as reference at the SOA-WC. (c) The recorded output (blue) and the expected output (red) with nonlinear referenced calculation. (d) the nonlinear transfer function of the SOA-WC, with correlation mapping of the input to output, with linear transformation (black) and nonlinear transformation (red).

in the blue line, and the expected inverted signal (calculated from Fig. 3.4*a*, with the linear transform as reference – Lin. Ref.) in the red line, resulting in an error of 0.14. By plotting the correlation map between the input and the output of the integrated SOA-WC detected at the PD, the optical nonlinear transfer function is illustrated in Fig. 3.4*d*, where the blue crosses are the data, the black line is the linear transform and the red line is the 3^{th}-order polynomial fitted nonlinear transform. The nonlinear function shape is mainly due to the contribution of the nonlinear response of the SOA used as wavelength converter when the booster works in transparency, with a current density of 1 kA/cm^2 and a weighting current density at 3 kA/cm^2 on average (linear regime). Then the same nonlinear function shape is utilized as a nonlinear reference (Nonlin. Ref.) to calculate the real expectation of the output, as shown in Fig. 3.4*c*, resulting in a smaller error of 0.08.

The pump input power of the wavelength converter can be tuned by increasing the current of the booster: A different level of pump input power provides a different transfer function shape. Fig. 3.5*a* presents the nonlinear functions when the current density of the booster is set at 0.5, 1.0, 1.5 and 2.0 kA/cm^2, with lines in blue, red, yellow and purple, respectively. The outputs near to level '-1' (for input level '1') tend to saturate when increasing the booster current, because of the nonlinearity changes due to the increased input probe power to the SOA-WC and because of the nonlinearity contributed from the booster SOA itself. This confirms that the nonlinear transfer function can be tailored by acting on the booster current. The SOA-WC based NL-TF shape as a function of the data rate is studied: Fig. 3.5*b* plots the nonlinear function when the input data rate is 2, 4, 6, 8 and 10 Gbit/s, with the blue, red, yellow, purple and green line, respectively, when the

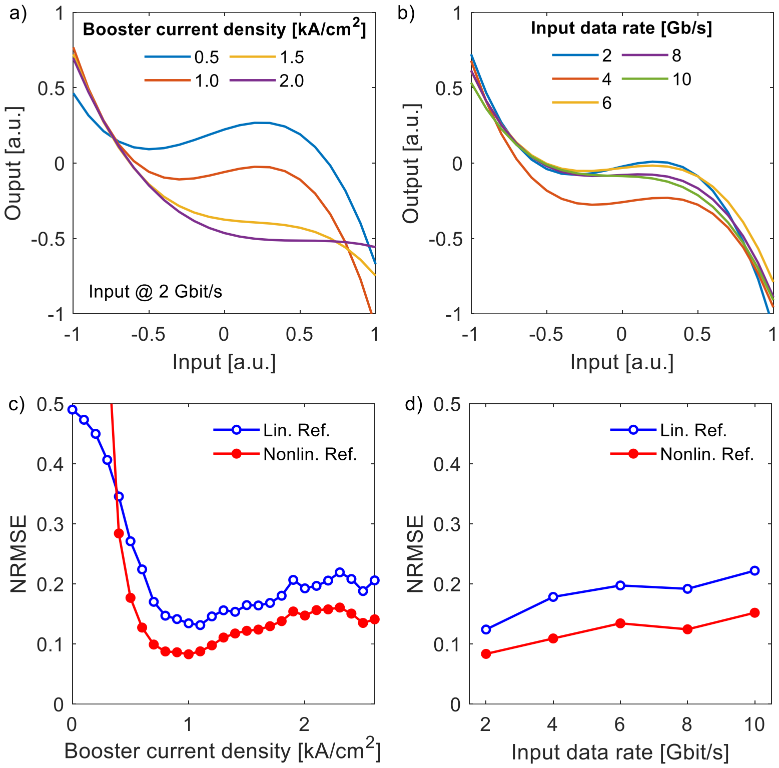


Fig. 3.5. (a) The nonlinear transfer function when the booster current density is set at different level. (b) The nonlinear transfer function when the input data rate changes from 2 to 10 Gbit/s. Error obtained at the neuron output when tuning the booster current density (c) and when tuning the input data rate (d), comparing to expectation calculated using linear transform as reference (blue) and nonlinear transform as reference (red).

booster is at 1 kA/cm^2. The shape of the nonlinear function changes only slightly when increasing the input data rate. These findings is translated into performance metrics of the optical nonlinear function by calculating the NRMSE with respect to different shapes of the nonlinear function. Fig. 3.5*c* plots the error variations of the output of the NL-TF when tuning the injection current density of the booster SOA from 0 to 2.5 kA/cm^2, with the blue line obtained when considering the neuron output as linear inverted output (with linear transform as reference), and the red line when considering the nonlinear transform as reference. The booster SOA is operated in the linear regime to minimize the nonlinearities introduced at the weighting element inputs, since the overall weighted addition operation is meant to be a linear operation. By changing the booster SOA current, one can find the optimal operation point, for minimized error induced by the nonlinear function, in this case corresponding to a current of 1 kA/cm^2 (Fig. 3.5*c*). The noticeable offset between the blue and red curves indicates that the nonlinearity of the SOA-WC has quite some effect on the error reduction. Fig. 3.5*d* plots the error variation when changing the data rate of the input from 2 to 10 Gbit/s, with the blue line showing the error related to the linear transform reference and the red line to nonlinear transform reference. In both cases, the error of the nonlinear function increases with the input data rate. Again, the nonlinear function improves accuracy, moving from 0.08 to 0.15 NRMSE when increasing the data rate up to 10 Gb/s. The deterioration in accuracy for higher data-rate is mainly due to the limited carrier lifetime of the integrated SOA-WC, which cannot fully follow the speed of the incoming optical signal. The offsets between the blue and red lines in both Fig. 3.5 *c* and *d* show that the use of the correct nonlinear transfer function reduces the error of up to 50%, compared to the case when using the SOA-WC with its simply linear response.

3.2.3. All-optical monolithically integrated neuron

The monolithic integration of the synaptic operation and the optical nonlinear function allow us to investigate the performance of the SOA-based all-optical neuron concept. The four-channel WDM PRBS-OOK input is coupled at the neuron input, with a data rate of up to 10 Gbit/s. The output is detected and compared to the calculated time trace with the NL-TF obtained following the procedure explained in session III-A. Fig. 3.6*a* plots the time traces as linear combination of the weighted input data, where the red line presents the recorded signal, and the

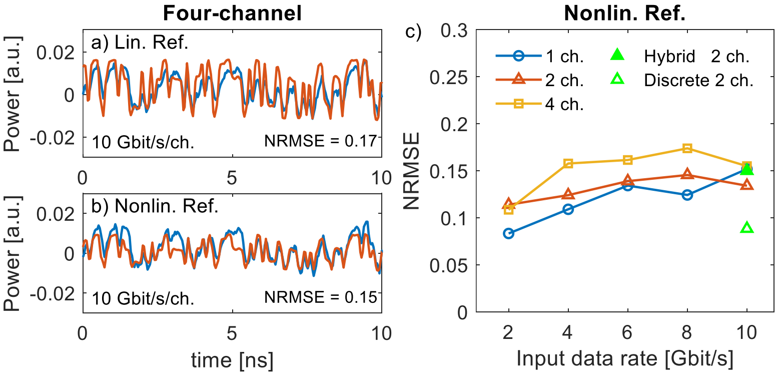


Fig. 3.6 The four-channel weighted addition recorded output (blue) comparing to the expected (red) (a) linear transformed reference, without NL-fitting, (b) nonlinear transformed reference with NL-fitting. (c) the error obtained at the neuron output for 1, 2, and 4 channel weighted addition (blue, red, and yellow), when tuning the data rate from 2 to 10 Gbit/s per channel, with green filled triangle denoting error of 2-channel hybrid integrated neuron and the unfilled triangle of 2-channel discrete neuron with optimized SOA-WC.

blue line is the expected linear combination of the weighted addition, without NL-fitting, resulting in an error of 0.17. Fig. 3.6*b* instead shows the output of the recorded output signal with nonlinear transform reference, where the blue line is the recorded signal and the red line the expectation, resulting in a smaller error of 0.15 - a 10% error reduction. The number of input channels and the data rate of the input signals are tuned to better analyse the performance of this all-optical neuron. Fig. 3.6*c* illustrates the error of the complete optical neuron output. The blue circle, red triangle, and yellow square symbols represent the errors of the all-optical neuron when the input channel changes from 1, 2 to 4 channels, respectively. In line with Fig. 3.6*d*, the curves show that the output error increases with the input data rate. Moreover, with the increase of the channel number, the error tends to increase as well. With more channel input to the SOA-WC, the nonlinearity at the SOA-WC is reduced as the increasing input probe power will push the cross-gain modulation regime towards a linear conversion regime. This means that an optimization of the operation regime of the wavelength converter is needed to help increase its nonlinearity, e.g., by tuning the power of the CW laser. Moreover, in Fig. 3.6*c*, a green-filled triangle is added to show an average error of 0.15, which is obtained when combining the integrated linear unit with a discrete nonlinear SOA wavelength converter, in section 2.5, with 10 Gbit/s per channel input and 2 channel weighted addition. This shows that the monolithically integrated all-optical neuron performs 10% better in terms of error introduction than the hybrid case under the same data rate condition. One reason for that can be that a discrete implementation generates additional noise due to the off-chip amplification. Finally, the integration of the tunable laser and SOA-WC also reduces the total power consumption as the external laser is not required, neither the additional off-chip amplifier. Further investigation shows that by using discrete SOA-WCs with optimized carrier dynamics, the multi-level conversion brings to a calculation error less than 0.09 [79], shown as green-unfilled triangle in Fig. 3.6*c*. In the next section, the simulation of an all-optical multi-layer neural network is presented by exploiting both the synaptic operation and the nonlinear function as realized and measured so far. A preliminary investigation of the AON is also reported in Appendix 3A for more insights into 1) how to set the optimal operation point of the no-linear SOA, as well as 2) what is the behaviour with data rate and number of inputs channels.

3.3. MNIST Classification with an SOA-based AONN

The combination of the linear neuron with the wavelength converter (section 3.2) eventually converts the multiple weighted wavelength inputs, after their addition, into one single wavelength which is the actual output of the complete neuron (yellow box in Fig. 3.1). In particular, the recorded transfer function of the integrated SOA-based wavelength converter, shown in Fig. 3.5*d*, has been evaluated in an analogue manner, with the power summation at the input of the SOA-WC being a multi-level signal. Therefore, the same transfer function will also work with multi-level WDM signals. This nonlinear function is then used to train the neural network on the computer via TensorFlow [67], while the pre-trained weighted matrix can be applied to the all-optical neural network to run inference and evaluate the accuracy.

The MNIST handwritten digit classification problem [80] is one of the benchmarking problems used for the performance appraisal of a neural network. The MNIST data set contains 60000 training samples and 10000 testing samples and includes 10 categories of digits from 0 to 9. In section 3.1, It shows that the linear synaptic operation of the SOA-based neuron can allow more than 64 channel inputs, with the introduction of negligible error. Here indeed the all-optical neural network is simulated with input layer neurons with 64 channel inputs each. To

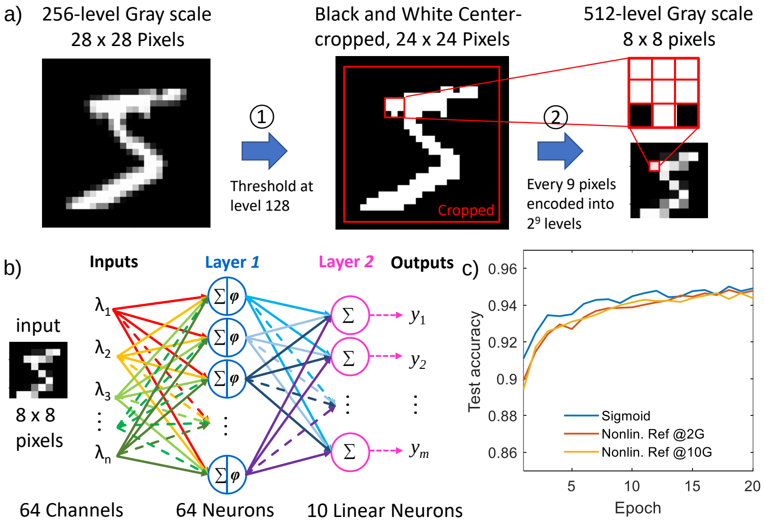


Fig. 3.7 (a) preprocessing of the input MNIST handwritten images from 728 pixels to 64 pixels. (b) The 2-layer neural network structure for classifying MNIST handwritten digits. (c) the test accuracy per epoch when training the 2-layer neural network with the sigmoid function (blue line) and the nonlinear functions when data rate is 2G Sample/s (red) and 10G Sample/s (yellow).

encode the input image into 64 channels by means of multi-level modulation with 9-bit resolution, the images in the dataset are pre-processed to reduce their resolution from 28×28 pixels to 8×8 pixels. Fig. 3.7*a* illustrates the data pre-processing for the input of the neural network (NN). The 256 level gray data is firstly converted into a black and white image with a threshold at level 128 and cropped into 24 by 24 pixels at the centre. The images are then converted to 8×8 pixels with every 3×3 pixels encoded into 512 grayscale levels, i.e. 9 bits-resolution. For solving this digit classification, a two-layer NN is structured as shown in Fig. 3.7*b*. On the 1st layer, there are 64 neurons where each of the weighted addition output is followed by the optical nonlinear function obtained in section 3.2.2, and on the 2nd (output) layer 10 linear neurons are used to represent the 10 digits, from 0 to 9. In the optical neural network (ONN) implementation, the inputs and the weights are usually normalized in order to ease the optical modulation and the dynamic weighting control. This is implemented in simulation by applying batch normalization and weight normalization. To train the NN for MNIST dataset classification, the ADAM optimizer is utilized due to its fast convergence [81], which makes the training process more efficient.

The two-layer structure is trained with the current third order polynomial nonlinear function without noise induction as a reference. The trained weighted matrix is then applied to the ONN model to investigate the performance of the optical network under error induction and contribution from the linear and the nonlinear units. Moreover, the same shallow neural network is benchmarked to the case using the sigmoid nonlinear function, with the same data as shown in Fig. 3.7*a*. The test accuracy is recorded after every update of the weighting matrices when training the neural network in TensorFlow. Fig. 3.7*c* presents the test accuracy as a function of the training epochs for different nonlinear functions: when the nonlinear function is the conventional mathematical sigmoid function (blue), when it corresponds to the transfer function observed at 2 GSample/s per channel (orange), and when it corresponds to the transfer function obtained at 10 GSample/s as input (yellow). Note that here the influence of the all-optical neuron impairments is not considered yet. The curves show that the NN is converging after 15 epochs of training and that all considered nonlinear functions yield similar final test accuracy of ~94.5% after training.

To take into account the error induced by the all-optical neuron, the distortion contributions is considered from the linear part of the neuron (described in section 3.1), and from the nonlinear part of the neuron (analysed in section 3.2). In particular, the distortions are included here as additive white gaussian noise, assuming the signal spontaneous emission beating noise dominates the contribution [82], which is added after the linear output and the nonlinear output. By tuning the

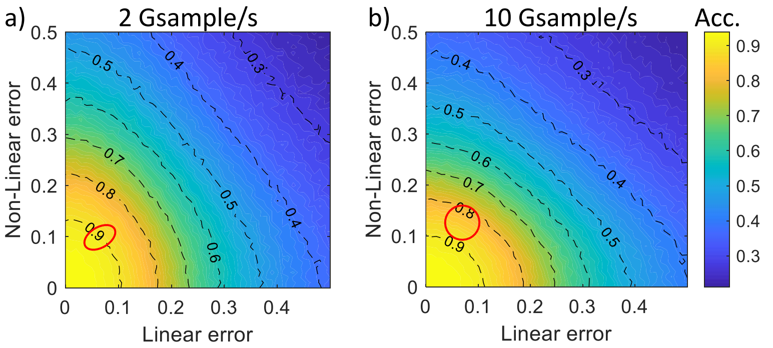


Fig. 3.8 The accuracy of the simulated noised ONN for nonlinear function recorded when input at (a) 2 GSample/s and (b) 10 GSample/s. With red circles presents the expected performance of using our all-optical neuron in the ONN.

standard deviation of the gaussian noise, the same error levels as the ones observed experimentally can be reproduced. The same inference is now run in the case impairments are induced in the optical neuron: Fig. 3.8*a* and *b* illustrate the colormap of the prediction accuracy (Acc.) as a function of the noise levels of both the linear and nonlinear functions of a neuron: these are scanned from 0 to 0.5, for both 2GSample/s and 10 Gsample/s input per channel, respectively. The accuracy in both cases obviously decreases when increasing the error at the output of both linear and nonlinear units. The red line shapes in Fig. 3.8*a* and *b* show the expected accuracy that the AONN system will have for a measured error level ranging from 0.05 to 0.10 for the linear operation (according to the error induced with 64 channel inputs, discussed in section 3.1) and for the nonlinear errors ranging between 0.08 and 0.11 for 2 Gbit/s input, and between 0.10 and 0.15 for 10 Gbit/s input, respectively, as recorded during the experiments. For these same areas, an accuracy degradation of 2-8% and 5-15% for 2 GSample/s and 10 GSample/s input, respectively, is obtained, compared to the trained accuracy of 94.5%. The elliptical shape in Fig. 3.8*a* is due to a different deviation of the gaussian noise distribution on the linear and nonlinear unit, while the circular shape in Fig. 3.8*b*, is due to a more uniform variation for both units. These suggest that with 10 Gsample/s input, the 2-layer all-optical engine, including 64 neurons in the first layer, with 64 synapses per neuron and 10 neurons at the 2nd layer fully connected, can perform 4.7×10^{13} MAC/s, which provides circa 2.5 times faster computation than state-of-the-art GPUs [83], and the same order as the TPU [23], considering only 5% best-case accuracy degradation and 10 GHz speed nonlinear processing, which is not available in GPUs and TPUs. Training the AONN with

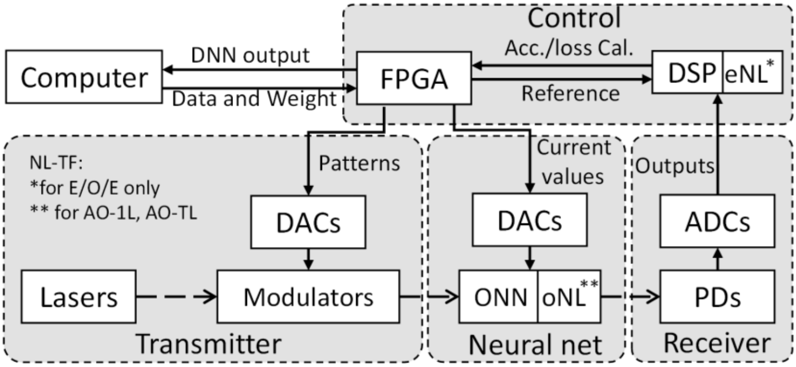


Fig. 3.9 Schematic of the considered end-to-end optical DNN system.

the addition of the estimated distortion from the linear and the nonlinear unit is expected to reduce the influence of the noise and preserve the high prediction accuracy of the NN using wavelength converter as nonlinear function, instead of the conventional sigmoid function. In future, it is expected that the scaling to 64 input neurons in our network system can be realized by interfacing the chip with high-speed state-of-the-art transceiver modules [84] or with co-packaged optics [85] in a multi-chip package.

3.4. System Energy Consumption Analysis

In this section, the power consumption is estimated on the end-to-end (digital-to-optical-to-digital) system enabling the implementation of the optical neural network. Fig. 3.9 shows the schema of the complete ONN system, which includes the transmitter, the optical chip, the receiver, the digital signal processor, and the control unit. The system overall is controlled by the control unit (Ctrl), which is interfaced with the computer and includes a field-programmable gate array (FPGA) and a digital signal processor (DSP). Here an FPGA is employed for the sake of fast development and reconfiguration flexibility [86],[87]. However, application specific integrated circuits (ASICs) can also be used to reduce the power consumption even further [88]. To analyse the effective power consumption of the ONN, all the components in the system should be taken into account. The transmitter (Tx) includes lasers, modulators and DACs, which are used to drive the modulators. The ONN includes the ONN chip and its control DACs and drivers for the weighting. The receiver (Rx) consists of photodetectors and the corresponding ADCs.

The energy consumption of the system is analysed by considering different operation modes of the ONN within the end-to-end system. Here one can consider three scenarios: (1) E/O/E with one linear layer (E/O/E), (2) All-Optical with one complete layer (AO-1L), (3) All-Optical with multiple layers (AO-TL), where *T* stands for *T* layers. In the E/O/E approach, the optical chip is used to calculate linear matrix multiplication, while the nonlinear function is realized on the DSP, with the data received at the PDs. In the all-optical approaches, the nonlinear function is co-integrated with the linear optical neuron, and the output (at each layer output for the AO-1L case, or at the end of the complete multiple-layer NN for the AO-TL case) is obtained via linear PDs. A more general ONN with N-inputs *M*-outputs and *T*-layers is now analysed, including the end-to-end system performance, for these three different operation modes. The operations executed by the ONN systems are different for these cases, depending on if a single layer or multiple layers are implemented. For the single-layer implementations, as in cases (1) and (2), the DNN needs to be decomposed into layers and analysed layer by layer, which is not necessary in case (3) for the same network implementation.

For the inference of a trained DNN, the data and weight matrix are loaded to the FPGA via the interface with a computer. The FPGA generates the electrical patterns as well as the weight control currents, which feed to the modulator DACs and the weight DACs and drivers, respectively, as shown in Fig. 3.9. The electrical patterns are imprinted on the laser beams and sent to the optical neural network chip. The chip is controlled with the analogue currents coming from the respective DACs and amplified at the drivers, with which the matrix multiplications are calculated. For the E/O/E case, the detected linear output is converted into digital signals by the ADCs, then the DSP unit processes the signals executing the nonlinear transfer function. The outputs are then sent back to the FPGA, which generates the patterns for the next layer. The next layer follows the same procedure. At the output layer, the outputs of the last layer nonlinear functions will be further processed by the FPGA and compared with the reference labels to provide the final prediction which is then passed to the computer. Therefore, the power consumption of the E/O/E single-layer system can be calculated as:

$$\begin{aligned} P_{E/O/E} = {} & N \times P_{Tx} + (N \times M) \times P_w \\ & + M \times (P_{Rx} + P_{eNL} + P_{ctrl}), \end{aligned} \tag{3.1}$$

where P_{Tx} is the power of transmitter per channel, P_w is the power for each weighting, including the power of DAC and the current driver for the ONN, P_{Rx} the power of receiver, P_{eNL} the power for the electrical nonlinear function and P_{ctrl} the power of the control.

For the all-optical single layer case AO-1L, the procedure is similar to the E/O/E case, with the only difference that the nonlinear function is co-integrated on the optical chip. Therefore, the DSP does not carry out the nonlinear function calculation and only calculates the final accuracy at the output layer. Hence, the power of the AO-1L system can be calculated as:

$$\begin{aligned} P_{AO-1L} = {} & N \times P_{Tx} + (N \times M) \times P_w \\ & + M \times (P_{oNL} + P_{Rx} + P_{ctrl}), \end{aligned} \tag{3.2}$$

where P_{oNL} is the power of the photonic nonlinear function.

Finally, for the all-optical T layer case AO-TL, the FPGA and DSP are not required to process and update the inputs and weights for the next layer but the DSP will calculate the loss and accuracy based on the final outputs and the reference labels. Therefore, the power consumption of the AO-TL system can be calculated as:

$$\begin{aligned} P_{AO-TL} = {} & N \times P_{Tx} + (N \times M \times T) \times P_w \\ & + M \times T \times P_{oNL} + M \times (P_d + P_{ctrl}). \end{aligned} \tag{3.3}$$

TABLE 3.1 Components in the optical neural network system

Components		O/E/O	AO-1L	AO-TL	Unit P (mW)	Ref.
Tx	Laser	N	N	N	150	[89]
	Mod.	N	N	N	20	[90]
	DACs	N	N	N	25	[91]
ONN	Weight Elements	N×M	N×M	N×M×T	30	S2.2
	DACs	N×M	N×M	N×M×T	25	[91]
	o-NL	-	M	M	150	S3.2
Rx	PDs	M	M	M	5	[92]
	ADCs	M	M	M	25	[91]
	e-NL	M	-	-	200	[87]
Ctrl	Accuracy Cal. Unit	M	M	M	200	[87]
	FPGA	M	M	M	200	

The required number of components of the three different scenarios and the power values used in the system power analysis are listed in TABLE 3.1. These values are considered when using state-of-the-art components that fit into the scheme of the SOA-based all-optical neural network structure as described in section 3.1.

Considering the delays related to all the components, the total time for the E/O/E system to execute one epoch can be specified as:

$$\begin{aligned} t_{E/O/E} = T \times (&S_N/f_{Tx} + 1/f_{Tx} + t_{Tx} + t_{oLin} \\ &+ 1/f_{Rx} + t_{Rx} + S_N/f_{eIO} + t_{eNL} \\ &+ t_{FPGA} + t_{e-inter}) + t_{acc}, \end{aligned} \tag{3.4}$$

for an AO-1L single layer system is calculated as:

$$\begin{aligned} t_{AO-1L} = T \times (&S_N/f_{Tx} + 1/f_{Tx} + t_{Tx} + t_{oLin} \\ &+ t_{oNL} + 1/f_{Rx} + t_{Rx} + S_N/f_{eIO} \\ &+ t_{FPGA} + t_{e-inter}) + t_{acc}, \end{aligned} \tag{3.5}$$

and for an AO-TL multi-layer system is:

$$\begin{aligned} t_{AO-TL} = &S_N/f_{Tx} + 1/f_{Tx} + t_{Tx} + T \times (t_{oLin} + t_{oNL}) \\ &+ 1/f_{Rx} + t_{Rx} + S_N/f_{eIO} \\ &+ t_{FPGA} + t_{e-inter} + t_{acc}, \end{aligned} \tag{3.6}$$

where S_N is the number of samples per epoch at the input of each layer, and f_{Tx}, f_{Rx} are the speed of the transmitter and receiver, respectively, t_{Tx}, t_{oLin}, t_{oNL}, t_{Rx}, t_{eNL} *and* $t_{e\text{-}inter}$ are the time delay from the transmitter, the optical linear unit, the optical nonlinear unit, the receiver, the electrical nonlinear function and electrical interconnection, respectively, t_{FPGA} is the computational time for the FPGA to generate the patterns and the current values for the weights and t_{acc} is the computational time of the DSP for the accuracy calculation. The average total energy consumption for epoch can be expressed as $E_{syst} = P_{syst} \cdot t_{syst}$, where E_{syst} is the total energy consumptions for the whole neural network system per epoch, t_{syst} is the time for computing one epoch of samples and P_{syst} is the total power of the end-to-end system, all calculated respectively for the 3 operational system cases.

The energy consumption for the optical MAC operation, i.e. the synaptic operation, depends on the number of controlled elements which provide the weights,

TABLE 3.2 Computing time of the components in the system

symbol	description	value	Unit
N	Max. Synapsis number in a neuron	64	
M	Max. Neuron number per layer	64	
T	Max. Layer number	10	
S_N	Input sample number	10^4	
$f_{Tx/Rx}$	Speed of optical transmitter/receiver	10	GHz
t_{Tx}	Time delay, transmitter	5	ps
t_{oLin}	Time delay, optical linear unit	10	ps
t_{oNL}	Time delay, optical nonlinear unit	20	ps
t_{Rx}	Time delay, receiver	2	ns
t_{eNL}	Time delay, electrical NL unit	3	ns
$t_{e\text{-}inter}$	Time delay, electrical connection	100	ns
f_{eIO}	Speed, I/O connection, FPGA	10	GHz
t_{FPGA}	Time, signal processing of FPGA	3	ns
t_{acc}	Time, acc./loss calculation of FPGA	6	ns

if only the optical engine is considered. Here the same weighting elements is considered, i.e. the SOAs, for which the power is 30 mW on average per weight, excluding the DACs. Therefore, for an operational input data rate of 10 GHz, the resulting power consumption for one MAC is 3.0 pJ/MAC and 5.5 pJ/MAC if the weight DACs is included. However, this estimation misses the contribution of the transceiver, the overall electrical controller, the receiver and the off-chip computations. Therefore, the end-to-end system power and the total computational time should be considered to obtain the real performance metrics of the optical neural network. For an N-input M-neuron T-layer DNN, the total number of MAC operations is $S_N \times M \times N \times T$. Hence, the effective energy consumption – *effective* as it includes the end-to-end system overall contribution – per MAC operation is the total power of the specific end-to-end system times the total time to execute one epoch, over the total number of MAC operations:

$$E_{MAC-eff} = P_{syst} \times t_{syst} / (S_N \times M \times N \times T)\,. \tag{3.7}$$

The delays and computational speed for different components are listed in TABLE 3.2. The values used in the calculations are considered based on off-the-shelf components. In particular, all optical delays are obtained from the actual path length, while all the electrical delays are related to the processing clock time of the off the-shelf electronics.

Firstly, the size scaling of the optical neural network is investigated. As mentioned in section 3.2, the network is considered to be scaling up with *M*=*N*, i.e., this energy analysis is done with respect to a quadratic scaling of the network. When the increasing number of neurons *M*, the splitting loss will increase. As a consequence, it compensates these losses with additional laser power by increasing N, the input channel number. From Table 3.2, it is clear that the largest DNN that under investigation is an *M*×*N*×*T* DNN with a maximum number of 64 input ×64 neuron/layer × 10 layers. Fig. 3.10 illustrates the $\mathrm{E}_{\text{MAC-eff}}$ obtained from the

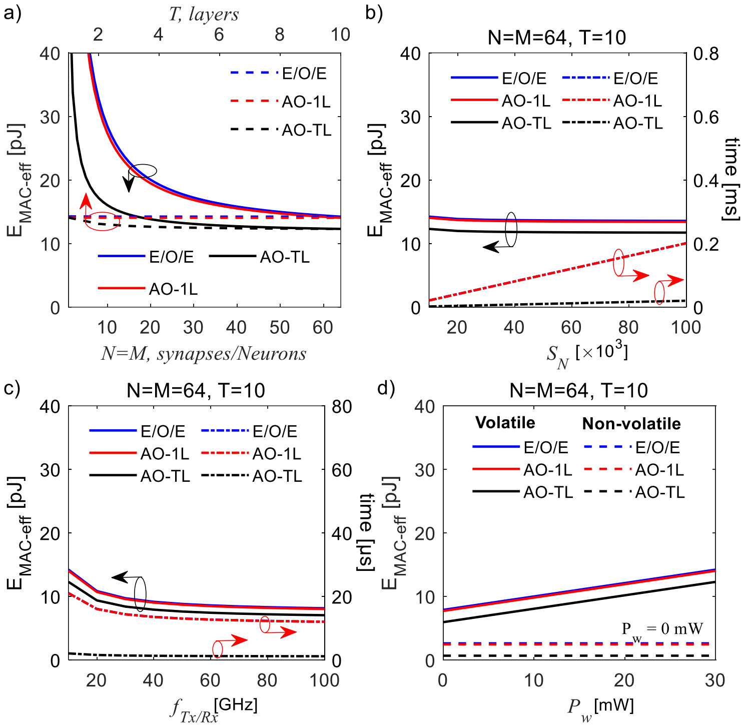


Fig. 3.10 The system energy consumption per MAC, $\mathrm{E}_{\text{MAC-eff}}$. (a) the $\mathrm{E}_{\text{MAC-eff}}$ obtained when changes the synapses number N (solid), and layer number T (dashed); (b)the $\mathrm{E}_{\text{MAC-eff}}$ (solid) and total computing time (dash-dot) vs. input sample numbers S_{N}; (c)the $\mathrm{E}_{\text{MAC-eff}}$ (solid) and total computing time (dash-dot) vs. speed of transceiver; (d)the $\mathrm{E}_{\text{MAC-eff}}$ calculated (solid) and total computing time (dash-dot) when changing the power of weighting elements P_{w}; for the E/O/E (blue), AO-1L (red), and AO-TL(black) neural network.

equations (3.1)-(3.7) for different system modes of operation and looking at different parameters.

Fig. 3.10*a* illustrates the energy consumption per MAC operation when increasing the number of synapses per neuron, N, with layer number T=10 (solid lines) and increasing the layer numbers T when fixing M=N=64 (dashed lines). The $E_{MAC\text{-}eff}$ for the multi-layer DNN is inversely proportional to the number of synapses for all the cases. The $E_{MAC\text{-}eff}$ for E/O/E (in blue) and AO-1L (in red) are very close, as in both cases the FPGA and transmitter for the signal processing and pattern regeneration, respectively, notably increase power consumption as well as computing time after each layer. The $E_{MAC\text{-}eff}$ tends gradually to the asymptotic value of 14 pJ/MAC. The lower limit of energy consumption is set by the power consumption at the transmitter side and at the weighting elements (this power relates to the weight unit power, therefore it does not depend on the synapses number). For the AO-TL neural network system, avoiding the electronics to optics to electronics conversions when moving layer by layer, the computing time gets reduced considerably: The rate of change of the $E_{MAC\text{-}eff}$ is faster than for the E/O/E and AO-1L cases, and reaches 12 pJ/MAC for a number of 64 synapses/neuron. If the FPGA was replaced with an ASIC with optimized designs to reduce the power consumption, the effective energy consumption would have not been changed dramatically since, in these particular large-scale network systems, the elements for the control of the weight represent the main contribution. Always in Fig. 3.10*a*, one can observe that the number of synapses per neuron in the system with single layer implementations should be greater than 20 for case (1), and greater than 18 for case (2), in order to guarantee an $E_{MAC\text{-}eff}$ down to 20 pJ/MAC, while for the AO-TL neural network this value is only 6. On the other hand, the dashed lines plot the $E_{MAC\text{-}eff}$ with respect to the number of layers when the synapses number N is 64. The $E_{MAC\text{-}eff}$ is in general very flat for all 3 cases. The difference among the single layer cases, E/O/E (1) and AO-1L (2), with the multi-layer case, AO-TL (3) is set by the synapse number: a bigger difference is expected for a smaller synapse number.

All the graphs in Fig. 3.10*a* tend to an asymptotic value because the lower limit is bound to the energy consumption on each synapse control component for M=N>64 and T>10. Hence, all the other investigations are carried out for M=N=64 and T=10, while changing other parameters like the input sample number S_N, the speed of the transceivers $f_{Tx/Rx}$ and the power of the weighting elements P_w. Fig. 3.10*b* presents the $E_{MAC\text{-}eff}$ and the total computing time when changing input sample numbers. The power efficiency only slightly decreases with varying the input sample numbers from 10k to 100k (solid lines) for two reasons: the $E_{MAC\text{-}eff}$ is calculated on each MAC operation of each sample and the total processing time for computing (dashed lines) increases linearly from 20 to 200 μs for the single-layer E/O/E and AO-1L neural network. The computing time for the AO-TL neural net case, instead, is at least 10 times faster. On the other hand, Fig. 3.10*c* shows the $E_{MAC\text{-}eff}$ as a function of the transmitter and receiver operation frequency. The energy consumption can be decreased 5 pJ for all the cases, when increasing the speed of the transmitter and receiver from 10 to 100 GHz, due to the reduction of the total computing time. Improvements of the SOA performance are though needed to enable high-speed all-optical signal processing: This is considered possible when exploiting concepts like quantum dot SOAs [93] or SOAs with carrier reservoir layer [94], for which carrier recovery times down to 0.5-10 ps have been demonstrated, which can facilitate operation bandwidth up to 100 GHz.

Finally, by tuning the power of the biased weighting elements, the energy consumption for 64-input 64-neuron 10-layer implementation with a transceiver speed of 10 GHz can be evaluated. Fig. 3.10*d* illustrates the resulting $E_{MAC\text{-}eff}$ when changing the power of the weighting elements from 0 to 30 mW (solid lines). The energy consumption per MAC rises linearly with the weight power from 8 to 14 pJ for the single layer cases, and from 6 to 12 pJ for the AO-TL DNN, so that the use of an all-optical multi-layer network gives a 14% improvement in effective energy consumption per MAC, with respect to E/O/E and the AO-1L implementations. In addition, the dashed lines show the $E_{MAC\text{-}eff}$ for the case when non-volatile weighting element is used, such as phase change materials [51]: For single-layer cases, the power consumption is 2.4 pJ/MAC, while for the AO-TL an energy consumption as low as 0.7 pJ/MAC is calculated. This energy is non-zero because of the transceiver and the post-processing on the FPGA, as shown in equations (1)-(3) (setting P_w=0 and P_{DAC}=0). This result suggests that the current control of the weighting elements contributes 5.3 pJ/MAC more for all the cases and that the SOA weighting consumes 6.3 pJ/MAC (obtained subtracting the energy consumption at 0 mW from the energy consumption at 30 mW). When substituting volatile and current biased elements with non-volatile elements in the AO-10L neural network, it shows up to 94% energy saving for each MAC operation. In any case, the energy consumption for AO-TL neural network outperforms single-layer neural network system implementations.

3.5. Conclusion

The performance of an all-optical neural network structure is analysed with WDM connectivity and SOA-based all-optical neurons. The linear neural network can be easily scaled as a function of WDM signals for multi-synapsis neurons: the linear processing unit can scale up to 64 synapsis with a large input dynamic range under neglectable error introduction. The monolithically integrated all-optical neuron is fully integrated and experimentally demonstrated exploiting an SOA XGM WC-based optical nonlinear function. The performance of the integrated all-optical neuron is 10% better than the hybrid case. The all-optical neural network is simulated with noise induction for benchmarking the inference of the noisy DNN with MNIST handwritten digits classification. The results show that when working with 10 Giga sample/s inputs, the computing speed is about 20 times faster than the state-of-the-art electronic GPU, while guaranteeing a best case accuracy decrease of accuracy of only 5%. Note that this investigation is done considering only a static noise modelling, where the noise is added layer by layer but a truly noise modelling is not considered, which is instead done in Chapter 4 to predict the ultimate scalability.

Further, an FPGA controlling and DSP unit with DACs and ADCs and transmitters and receivers, is introduced to emulate the complete end-to-end system. The energy consumption is analysed on a system level and take into account all components utilized in the system when an N-input M-neuron T-layer DNN is implemented. The calculation results show that the energy per MAC operation for an all-optical connected multi-layer DNN outperforms the signal-layer DNN system. The energy efficiency is constrained by the speed and power consumption on the electronic side when increasing the number of synapses/neuron, but still it performs about 4 times better than state-of-the-art GPUs at the server level [83], excluding the energy for the cooling.

Appendix 3.A. Preliminary assessment of Monolithically Integrated AON

In this appendix, the preliminary assessment of the AON is demonstrated. The feed-forward neural network considered consists of a network of concatenated layers of neurons, where information travels from the left to the right side. The conceived all-optical deep neural network architecture is showed in Fig. 3A.1*a*. Each neuron (as in the grey box) receives at its input a number of different wavelength signals and it gives as output a signal coded into yet another wavelength. The output of this neuron, together with the outputs at the other neurons of the same layer, are then sent to the next layer of neurons for further processing, and so on and so forth.

In this experiment, the full photonic integrated neuron is realized by using a combination of arrayed waveguide gratings (AWGs) and semiconductor optical amplifiers (SOAs) technology. Fig. 3A.1*b* shows the schematic of the implemented photonic neuron. The neuron processes *N* wavelength division multiplexed (WDM) signals. After pre-amplification, the signals enter a de-multiplexer which allows access to the individual channels. These are weighted by using the gain variation of multiple SOAs, then a multiplexer combines back all the

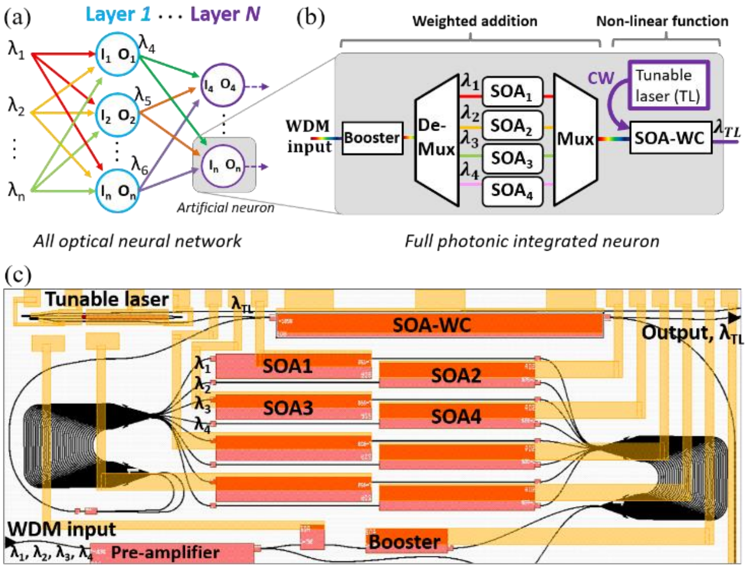


Fig. 3A.1 (a) Representation of the all-optical deep neural network, with white circles being the neurons. (b) Scheme of the implemented monolithically integrated neuron. (c) Mask details of the chip. WC=wavelength converter. CW= continuous wave.

wavelength signals. The weighted WDM signals undergo the SOA-based wavelength converter (SOA-WC), which is employed as optical activation function. The tunable laser is co-integrated in the photonic chip. The SOA-WC converts all-optically the multi-wavelength signal total power into one single wavelength, which represents the neuron output. Although in this first demonstration the SOA-WC provides an inverted signal at its output, an SOA-MZI scheme can be used in the future as non-inverted WC [56].

Fig. 3A.1*c* illustrates the mask details of the complete on-chip integrated neuron which includes the weighted addition circuitry and the SOA-WC. The input WDM signals are first amplified by a pre-amplifier and a booster SOA, and then sent to the weighted addition circuitry, composed of two 1×8 AWGs, and eight 950 µm long SOAs. The first 1×8 AWG operates as wavelength de-multiplexer, the second one operates as wavelength multiplexer. Both AWGs are designed with free spectral range (FSR) of 2.4 nm. Setting the injection current of the 8 SOAs determines the weight of each of the input data. The WC is based on cross-gain modulation [78], with a 2 mm long SOA and a tunable coupled-cavity laser (TL). The TL, centered at 1549.0 nm, provides 0 dBm optical power. Each of its cavities incorporates a gain section for laser actuation and a phase-tuning section for the laser tunability [76]. The chip employs a combination of SOAs at its input side to increase the SOA linear range. The broadband operation of the SOA enables the weighting of any wavelength in the C-band. The 4.6×2 mm^2 fabricated photonic integrated chip has been processed via an InGaAsP/InP multi-project wafer run.

3A.1 Experimental setup

The experimental set up to assess the operation of the monolithically integrated photonic neuron and a micrograph image of the chip are shown in Fig. 3A.2. Four input data at λ_1=1544.0 nm, λ_2=1546.0 nm, λ_3=1551.0 nm and λ_4=1554.0 nm are multiplexed and modulated by an optical modulator driven by an NRZ PRBS signal produced by an arbitrary waveform generator at 2 Gb/s. The modulated WDM signals are amplified by an EDFA, de-correlated and synchronized by using varied-length fibers and tunable time delays. Polarization control is performed on each channel to maximize SOA gain on chip. The WDM signals are then input to the PIC port 1 via a lensed fiber. The neuron optical output is filtered by a 0.4 nm bandpass filter, centered at the tunable laser central wavelength and detected by an AC coupled PD (DSC-R402APD) and a Digital Phosphor Oscilloscope (DPO, Tektronix), where the time traces are digitized and recorded. The input SOA and

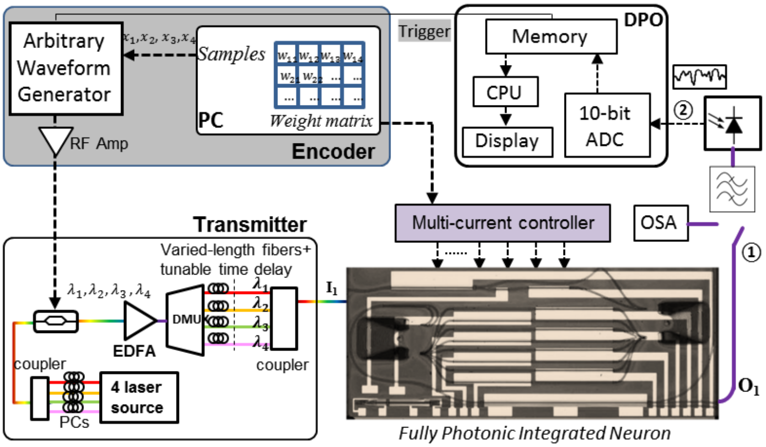


Fig. 3A.2 Experimental setup for the assessment of the photonic integrated neuron. OSA = optical spectrum analyser.

the WC SOA are set at 90 mA and 120 mA, respectively. For this first demonstration, 4 SOAs (SOA1 to SOA4 in Fig. 3A.1*c*) are biased with different currents controlled via a multi-current controller (Thorlabs), in order to assign the gain value which acts as a *weight factor* to the corresponding input data. But eight inputs are also possible.

The characterization of the all-optical photonic integrated neuron is carried out to calculate its processing accuracy. This is done by calculating the normalized root mean square error (NRMSE), i.e. the discrepancy between the measured and expected results at the neuron output, which is a number between 0 and 1. The lower the NRMSE, the better the accuracy.

3A.2 Experimental results

Firstly, one input channel is inserted to the neuron (Fig. 3A.3*a*) to optimize the wavelength converter operation as an integrator and non-linear function. Fig. 3A.3*b* shows the output of the photonic neuron after the SOA-WC: The measured output (blue line) is good matching to the expected inverted signal (red line), with an NRMSE of 0.13 (87% of accuracy). This experimental accuracy can be improved by increasing the optical input power budget and removing un-necessary sources of noise, but it is already comparable to the accuracy of a digital computer with 64-bit resolution in neural network implementation.

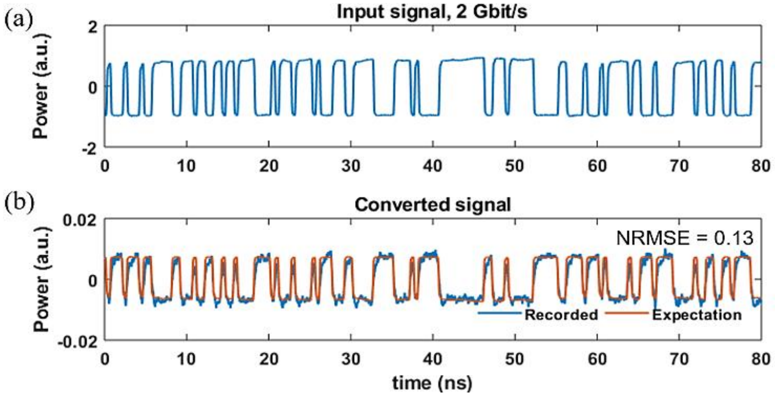


Fig. 3A.3 (a) Single input channel. (b) Measured (blue) and expected (red) neuron output. An error of 0.13 is calculated.

The input power of the modulated optical data (probe signal) to the wavelength converter is critical for the cross-gain modulation: There is an optimum value of probe optical power for the system performance. Therefore a power optimization procedure is investigated. The booster SOA is tuned in a 26 mA range to optimize the total optical power of the input modulated data fed into the SOA-WC. Fig. 3A.4*a* shows the error variation obtained while tuning the injection current at the booster SOA. There are two regions where the final accuracy is visibly degraded, which distinguishes an optimal operational regime where the NRMSE reaches its minimum. Fig. 3A.4*b* shows the measured output signal (blue line) and the

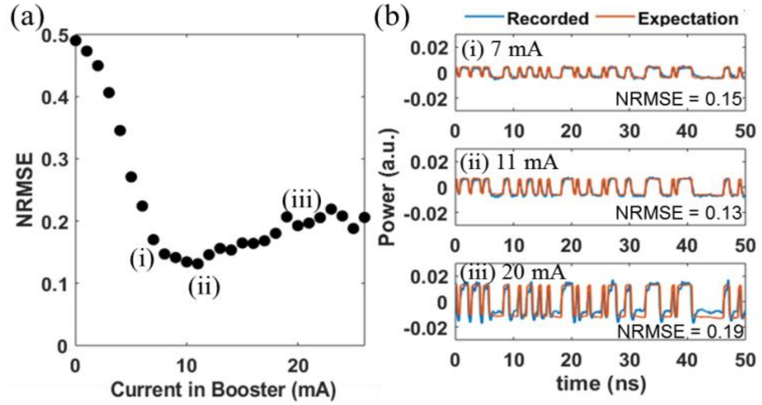


Fig. 3A.4 Optimization of the probe power to the wavelength converter SOA by tuning the current injection in the booster. (a) NRMSE vs. Current in the booster. (b) Time traces at injection currents of: (i) 7mA, (ii) 11mA and (iii) 20 mA.

expected signal (red line) for the cases where the injection current of the booster SOA is set to 7 mA, 11 mA and 20 mA, resulting into an error of 0.15, 0.13, 0.19, respectively. The lower peak-to-peak values obtained for low current values (Fig. 3A.4*b* (i)), i.e. lower probe optical powers, are caused by the low carrier-to-noise ratio. At higher power level (Fig. 3A.4*b* (iii)), an increase in the probe optical power increases its contribution to the gain saturation effect, thereby reducing the conversion efficiency. The optimized current is found to be 11 mA (Fig. 3A.4b (ii)).

Once the optimal input power to the wavelength converter is found, the full photonic neuron operation with 4 WDM input data is tested as shown in Fig. 3A.5*a* (total input power of -6.5 dBm). Fig. 3A.5*b* illustrates the optical spectrum at the chip output, where the peaks from left to right are at wavelength λ_1, λ_2, λ_{TL}, λ_3 and λ_4, (total unfiltered output power of -4.1 dBm) for a net on-chip gain of 11.4 dB. The different peak powers of each wavelength signal are due to the different weights assigned to each individual WDM signal by changing the gain of the SOA1 to 4. The measured photonic neuron output after being filtered and after the O/E conversion is shown in Fig. 3A.5*c* (blue line). The expected results are also reported (red line) in order to estimate the accuracy of the fully integrated photonic neuron. The calculated NRMSE of the neuron for 4 channel input is 0.11 (89% accuracy).

The error becomes slightly smaller as the total input signal power is increased, as the associated signal to noise ratio (SNR) is also improved. Furthermore, the

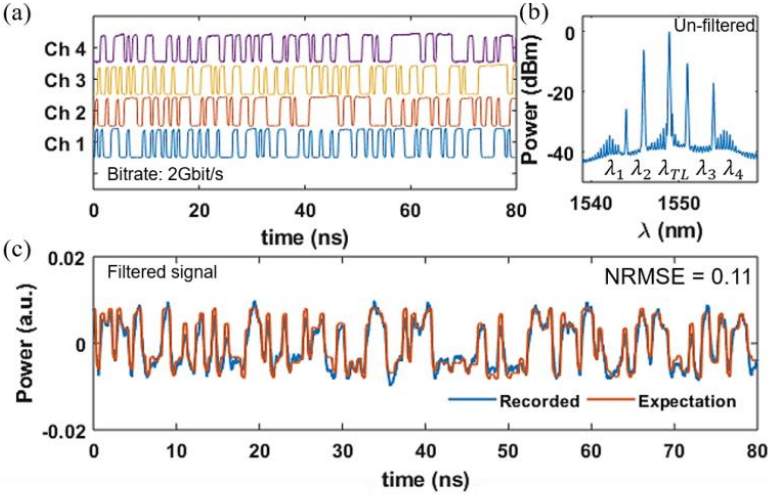


Fig. 3A.5 (a) The four input channels. (b) Un-filtered output spectrum. (c) Filtered and detected time traces at the neuron output for four channel WDM input processing.

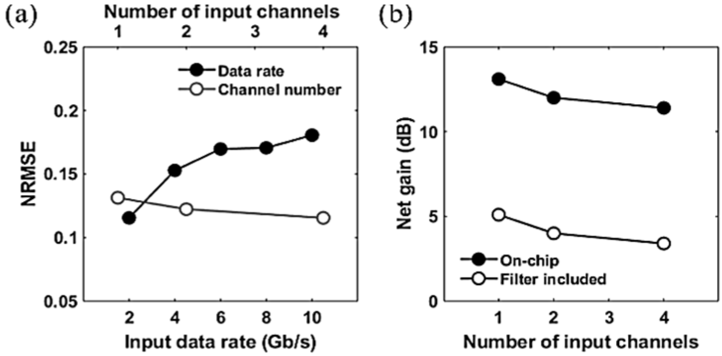


Fig. 3A.6 (a) Error evolution when increasing the input data-rate (filled circles) and the number of input channels (empty circles). (b) On-chip gain (filled circles), and if the external filter was co-integrated (empty circles) as a function of the number of input channels.

response of the optical neuron is measured when increasing the input signal data rate. In Fig. 3A.6*a*, the error gradually increases from 0.11 to 0.18 with the increase of the four input channel data rate from 2 Gb/s up to 10 Gb/s (filled symbols). This trend is in line with the carrier dynamics of both the booster SOA and WC SOA. From the measurement results, the '-1' levels appear to present overshooting, which is possibly due to the nonlinear effects at the booster SOA, as this effect increases while increasing the booster SOA injection current. The time traces also show smoother rising edge when the signal symbol changes, which is due to the carrier lifetime of the wavelength converter SOA. Fig. 3A.6*a* also includes the error variation (empty circles) when 1, 2 and 4 channels are input to the optical neuron. Fig. 3A.6*b* depicts the net on-chip gain (filled circles), and if the external filter was co-integrated (empty circles), as a function of the number of input channels. The net on-chip gain is calculated considering 11 dB fiber-to-chip total coupling losses. Both the final accuracy and the on-chip gain improve while increasing the number of input channels. A higher number of input channels has offered a higher probe optical power to the wavelength converter, resulting in improved conversion efficiency for the used current settings.

Chapter 4

Noise Evolution and Scalability Investigation of SOA-based AONN

In this chapter, the noise evolution and scalability of the proposed SOA-based AON is investigated as a function of multi-wavelengths and multi-layers by with noise modelling. The noise evolution along with the processing of the optical signal is of fundamental importance to evaluate the performance of the AON against the input noise and therefore to determine the scalability of the AONN in terms of multi-wavelength input and multi-layers. A noise model is developed to simulate the response of the AON, and an experimental emulation of the scaling of the AONN is demonstrated to analysis the capability of the AONN with the model. The noise modelling of the SOA-base all-optical neuron is explained in section 4.1. The experimental emulation of the AONN is demonstrated in section 4.2, the error evolution emulated layer after layer, and the network scalability is predicted in terms of its own height and depth.

4.1. Noise Modelling of All-optical Neuron

To understand the performance of the AONN at large scale, a noise model is needed to emulate the cascade of many layers, as well as the increasing number of input channels. A B2B measurement at the APD is necessary to determine its equivalent OSNR and to use it as a reference to estimate the layer OSNR.

The NRMSE at the APD can be estimated as (see Appendix 4.A):

$$NRMSE = \sqrt{N_e}/S, \tag{4.1}$$

where N_e is noise of the photodetector, and S is the detected signal span of the optical signal. By setting the input optical signal span power as constant, firstly an empirical relation between the spontaneous emission source noise and the error on the time traces is derived (see Appendix 4.A):

$$NRMSE = \sqrt{c_1 I_{sp}^2 + c_2 I_{sp} + c_3}, \tag{4.2}$$

where I_{sp} is the spontaneous emission noise at the APD, and c_1, c_2 and c_3 are the coefficients related to the spontaneous-spontaneous beating noise, signal-spontaneous beating noise and thermal noise over the fixed signal power, calculated from equation (4A.14) – (4A.16) in Appendix 4.A. One can now determine the noise level for a given NRMSE value using the inverse relation of equation (4.2), therefore estimating the equivalent OSNR at the neuron output. This reference NRMSE can be measured with B2B measurements.

To estimate the performance of the AON, the noise accumulation along with its propagation in the AON is considered. The AON includes 3 stages of SOAs: pre-amplifier, weighting SOAs and WC-SOA, each contributing to the noise build up along with the signal propagation through the neuron. The noise from the SOAs is obtained once given the inversion parameter n_{sp} and the single-pass signal gain G. Moreover, G is defined by the relation between input power and output single-pass signal gain (see equation (4B.4) in Appendix 4.B).

When analysing the nonlinear transformation with SOA, it is important to note that the input power to the WC-SOA is too weak to enable four-wave-mixing effect [95]. Since a carrier lifetime τ for the WC-SOA is estimated to be about 200 ps [71], the FWM effect is neglectable when detuning between the probe and pump channel is $\Delta f \gg 1/(2\pi\tau) \approx$ 1GHz. The experimental work in this paper is carried out with channel spacing of 400 GHz and simulations are implemented for 400 and 100 GHz, which is far greater than 1 GHz and results in a conjugate signal generated by FWM < -64 dBm.

To estimate the noise at the output of the WC-SOA, considering the contribution from the input signals and spontaneous noise at the WC-SOA input, as well as the ASE generated by the WC-SOA itself, the spontaneous emission density at the neuron output with WDM inputs is defined as [96]:

$$\rho_{ASE} = \sum \eta_i \rho_{sse,i} + \sum \tilde{\eta}_i \rho_{wc-ASE,i}\,, \tag{4.3}$$

with

$$\eta_i = |P_i/(P_T w_i) \cdot F(L)|, \tag{4.4}$$

$$\tilde{\eta}_i = \eta_i / G\,, \tag{4.5}$$

$$w_i = p_i / p_T\,, \tag{4.6}$$

$$p_T = \sum p_i, \quad P_T = \sum P_i\,, \tag{4.7}$$

where $\rho_{sse,i}$ and $\rho_{wc-ASE,i}$ are the spontaneous source emission (SSE) density at the input of WC-SOA and amplified spontaneous emission density from WC-SOA for the i-th input WDM channel. η_i, w_i, p_i, and P_i are conversion efficiency, weighting factor, small signal modulation and the averaged optical power for the i-th input channel, respectively. Note in equation (4.3), the CW laser channel (0-*th* channel) is included and $\eta_0 = \tilde{\eta}_0 = 1$. $F(L)$ is a function of P_T when the length

L of the SOA is fixed [97],[98] (see equation (4B.4) in Appendix 4.B). The saturated gain is assumed to be the same for all the input signals for simplified calculation with the dense WDM input signal within the gain bandwidth of the WC-SOA. The first term in equation (4.3) represents the noise conversion from the input noise and the second term shows the internal ASE contribution of the WC-SOA.

In equation (4.3), η_i, $\tilde{\eta}_i$, $\rho_{sse,i}$ and $\rho_{ASE,i}$ need to be given to calculate the output ASE. The conversion efficiency $\eta_i$and $\tilde{\eta}_i$can be determine by working out F(L) in equation (4.4), in which the unsaturated gain G_0, saturated power P_{sat} and normalized waveguide loss α' should be given. They can be found by measuring the SOA gain as a function of the input power and fitting the curve via the equation (4B.4).

The input SSE density $\rho_{sse,i}$ at the input of WC-SOA can be changed by tuning the neuron input OSNR, with the total gain provided from the pre-amplifier SOA (indicated with *pre*) and the weighting SOAs (indicated with *w*), with the total loss coming from the passive components such as splitters and AWGs. However, the ASE density from the pre-amplifier and the weighting-SOAs, $\rho_{pre/w-ASE}$,can be estimated by measuring the neuron output spectrum when the OSNR at the neuron input is maximum, in this experimental case OSNR = 55 dB, and then subtracting the amplification coming from the WC-SOA. Moreover, the $\rho_{pre/w-ASE}$ can be determined as a function of the input channels considering the gain and noise changes when tuning the input channel numbers and the relative input OSNR. Since the pre-amplifier SOAs and weighting SOAs are always working in the linear regime, the relative variation of gain and noise are supposed to be small. The noise density at the centre wavelength λ_c for the pre-/weight SOAs and WC-SOA are modelled by [99]:

$$\rho_{ASE,c} = h\nu_c\{n_{sp}[G-1] + b_{sp}(P_T/P_{sat})\ln(G_0)\}, \tag{4.8}$$

where h is the plank constant, ν_c is the optical frequency of centre wavelength of the SOA, n_{sp} is the inversion parameter for spontaneous emission, and b_{sp} is the inversion parameter related to the noise saturation. Note that these parameters may differ for pre-amplifier/weight SOAs and WC-SOAs due to different lengths, current densities, qualities, and properties of SOAs used in the experiment. They can be found by measuring the specific SOA noise versus input power. The WC-SOA output noise density of the *i*-th channel λ_i can be approximated by a second-order polynomial in logarithm [100]:

$$\log(\rho_{ASE,i}) = (a_1(\lambda_i-\lambda_c)^2 + a_2(\lambda_i-\lambda_c) + 1)\log(\rho_{ASE,c}), \tag{4.9}$$

where the maximum ASE is at the centre wavelength. Equation (4.8), (4.9) are the general noise models for SOAs, for respecting the ASE generation from pre-amplifier SOA, weight SOA, and WC-SOA. Hence, the ASE values for centre wavelength can be denoted as $\rho_{pre-ASE,c}$, $\rho_{w-ASE,c}$ and $\rho_{wc-ASE,c}$. And their corresponding i-th ASE density as $\rho_{pre-ASE,i}$, $\rho_{w-ASE,i}$ and $\rho_{wc-ASE,i}$. Equation (4.8) can be obtained by measuring the ASE from WC-SOA as a function of the neuron input power curve, while equation (4.9) can be obtained by measuring the output optical spectrum. Therefore, the spontaneous source emission noise at the WC-SOA input is:

$$\rho_{sse,i} = \rho_{sse0,i} L_c G_1 L_1 G_i L_2 + G_i L_2 \rho_{pre-ASE,i} + L_2 \rho_{w-ASE,i}, \qquad (4.10)$$

where $\rho_{sse0,i}$ are the original SSE noise at i-th channel input, L_c, L_1, and L_2 are the losses from the coupling, output passive loss of pre-amplifier SOA, and output passive loss of weight SOA, respectively, including all the lossy components on the path. G_1 and G_i are the gain provided by the pre-amplifier SOA and i-th weight SOA for individual weighting. Similarly, the input signal power to the WC-SOA is calculated as:

$$P_i = P_0 L_c G_1 L_1 G_i L_2, \qquad (4.11)$$

where P_0 is the input signal power per channel. Up to now, equation (4.2)-(4.7) can be used to obtain the ρ_{ASE} at the output. After filtering, the output noise power of the WC-SOA is:

$$P_{wc-SOA} = L_{OBPF} \rho_{ASE} B_o, \qquad (4.12)$$

where L_{OBPF} and B_o are the loss and 3dB bandwidth of the optical bandpass filter, respectively.

The NRSME can be estimated with the N_{wc-SOA} over the optical signal power, considering the conversion efficiency of the optical signal (see Appendix 4.C):

$$NRMSE = \sqrt{(c_1' I_{wc-SOA}^2 + c_2' I_{wc-SOA} + c_3')/\eta}, \qquad (4.13)$$

where $c_{1'}$, $c_{2'}$, and $c_{3'}$ are the co-efficiencies for the same terms defined in equation 3, with different values calculated in Appendix 4.C, from equation (4C.7) - (4C.9), and $\eta = \sum \eta_i$, given by equation (4C.3).

With equations (4.3) - (4.13), by tuning the input OSNR, the error evolution of the AONN can be determined. And with equation (4.2) we can evaluate the equivalent output OSNR. Note that the input OSNR and the output NRMSE are directly measurable data with the experimental setup described in the section 4.2. The parameters used in the simulations are determined from the measurement results in the SOAs and photonic integrated all-optical neuron characterisation. The parameters are then used to calculate the NRMSE as a function of the number of input channels and of the input OSNR (or number of layers) and is then compared to the error evolution from the measurements.

Then, the condition of the noise compression for the SOA-based AON can be estimated. The noise generation of the WC-SOA is depended on the input noise source level. With the OSNR for the input of the WC-SOA, P_{in}/P_{sse}, with an output OSNR, P_{out}/P_{ASE}, the input OSNR is:

$$OSNR_{in} = \frac{\sum P_i}{\sum P_{sse,i}} = \frac{\sum P_i}{B_o \sum \rho_{sse,i}}, \tag{4.14}$$

and the output OSNR is

$$OSNR_{out} = \frac{P_{out}}{P_{ASE}} = \frac{GP_{cw}}{\rho_{ASE} B_o}. \tag{4.15}$$

For a noise suppression:

$$\frac{OSNR_{in}}{OSNR_{out}} \leq 1. \tag{4.16}$$

In a simple case, when $\eta_i = \eta/M$, with $\bar{\rho}_{sse} = \frac{1}{M}\sum \rho_{sse,i}$, $\bar{P}_{in} = \frac{1}{M}\sum P_i$ and $\bar{\rho}_{ASE} = \frac{1}{M}\sum \rho_{ASE,i}$, following derivation in Appendix 4.D, it yields to,

$$OSNR_{in} \leq \frac{GP_{cw} - \eta \bar{P}_{in}}{\eta \bar{\rho}_{ASE}/G + G\,\rho_{sse,c} + \rho_{ASE,c}} \cdot \frac{1}{B_o}, \tag{4.17}$$

where P_{cw} is the power of CW laser, and $\bar{P}_{in}$ is the average power of the input signal, G is the saturated gain of the WC-SOA. The equation (4.17) shows that it is possible to find an input OSNR for noise suppression if the $GP_{cw} > \eta \bar{P}_{in}$, and $\bar{\rho}_{ASE}$, $\rho_{ASE,c}$ is defined by the noise generation from the WC-SOA as shown in equation (4.8) and (4.9), and the noise from CW laser $\rho_{sse,c}$ is either defined by equation (4.10) in this experiment, or by the SSE of a direct coupled CW-laser to the WC-SOA, as shown in Chapter 3. If the input OSNR satisfies the requirement in equation (4.17), the output noise is compressed due to the wavelength conversion and the amplification of the CW-laser power. Note $\eta_i = \eta/M$ is the result from equation (4.4) when input signals are equalized in power. The condition of equation (4.17) for un-equalized WC-SOA inputs can be obtained numerically with equations (4.3), (4.13) and (4.14).

4.2. Noise Evolution Emulation

For the emulation of the neural network scalability, the performance of the neuron is evaluated by measuring the optical signal output with changing the number of input channels while the performance of the neural network with the cascade of layers of neurons can be evaluated with changing the input noise. In particular, the number of input channels to the neurons defines the connection capability from all the neurons in the previous layer to the next layer. Therefore, determining the allowed number of input channels to a neuron is equivalent to the possible number of neurons per layer. This defines the height of the neural network. Moreover, for the layer scalability, the signal degradation is emulated, defined by the OSNR at the input and output of each layer. However, the output OSNR at the wavelength conversion is not measurable due to a non-observable in-band noise after conversion. For the sake of estimating the output OSNR, we determine the normalized root mean square error (NRMSE) obtained from the output time traces: with a given OSNR at the input, the neuron will show a level of error at the output, which can be reconducted to an equivalent level of OSNR at the output. If one can estimate the OSNR of the neuron output, this can be seen as the input OSNR to the next layer. By connecting the OSNR-error-OSNR relation of a neuron, we can emulation the scaling of the neural layers, which defines the depth of the neural networks.

To assess the error evolution of the optical signal and evaluate the performance of the SOA-based AONN, we use a monolithically integrated SOA-based all-optical neuron. Fig. 4.1*a* presents the structure of the all-optical neural network with WDM channel interconnections consisted of multiple AONs per layer and feed forwarded to deeper layers. The circles represent the AONs with optical linear and nonlinear units for weighed addition $\sum$ and activation function φ, respectively. The M-th neuron in the t-th layer (grey box) gives an output $y_{t,M} = \varphi(\sum w_i x_i)$, imprinted on $\lambda_{t,N}$, with w_i and x_i being the i-th weighting factor and input data, respectively. Fig. 4.1*b* sketches the schematic of the SOA-based AON, with WDM input and single channel output. The data is encoded as amplitudes in the optical carriers with different colours. The weight SOA is used to assign the weighting on the individual input, providing the gain to the input signal amplitude. Then the weighted WDM signal is multiplexed and sent to the NL-SOA. The summation of the weighted WDM signal is converted to a single output wavelength, via a nonlinear transformation, which serves as the input signal to the next layer of neurons.

Fig. 4.1*c* illustrates the experimental setup for assessing the performance of the SOA-based all-optical neuron (AON) and emulation of the scalability of the AONN. As shown in the grey box in Fig. 4.1*b*, the multiple inputs of the all-optical neuron are weighted and then summated at the SOA-WC and converted to a single wavelength, with optical signal propagating to the next layers. The signal degradation defines the maximum number of input wavelengths, therefore limiting the number of neurons that can be used simultaneously in one layer, for a certain error level induced. In this paper, we use up to 7 WDM inputs, matching the on-chip passband wavelength defined by the arrayed waveguide grating

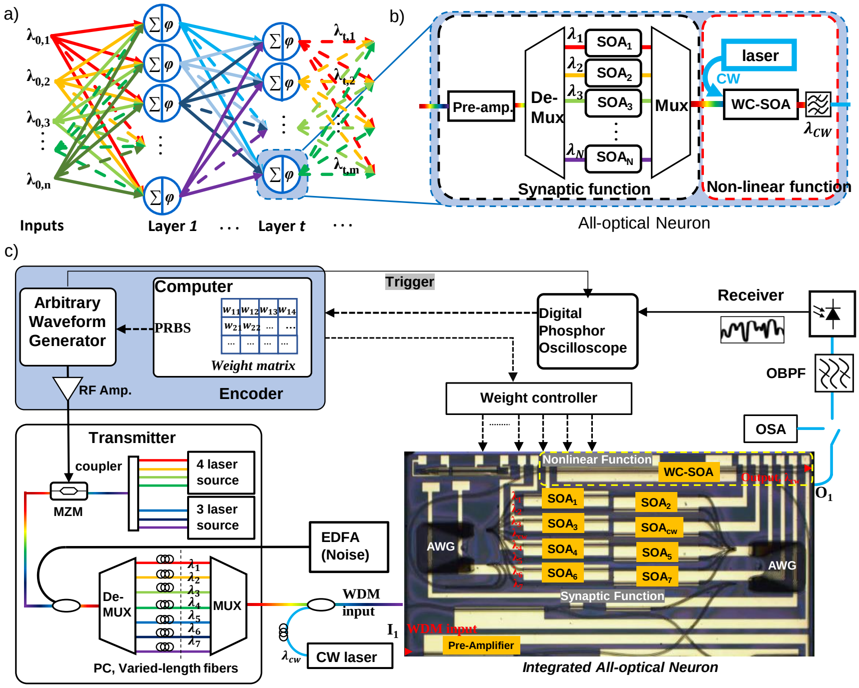


Fig. 4.1 (a) Structure of photonic deep neural network, (b) Schematic of SOA-based All-optical neuron, (c) Experimental setup for scalability emulation of all-optical neural network. PC: Polarization Controller, OSA: Optical Spectrum Analyser, OPBF: Optical Bandpass Filter.

channel separation, operated at 1540.3, 1542.5 1544.8, 1549.5, 1552.1, 1554.5, and 1556.8 nm. These are amplitude modulated with 10 Gbit/s PRBS signal, generated by an arbitrary waveform generator. Then the input is sent to the AON after decorrelation and multiplexing, with a power of -17.5 dBm per channel. In this work, an external continuous wave (CW) laser at 1546.72 nm is multiplexed to the WDM inputs to replace the tunable laser on-chip, in order to achieve a better wavelength conversion at the WC-SOA. Previous work has shown that the poor sideband suppression of the on-chip laser is limiting the quality of the converted output. However, this can be improved with a better laser cavity design for future use in the on-chip AON.

In the AON, the optical WDM input, pre-amplified with an SOA at 60 mA, is de-multiplexed and weighted by the SOAs which are operating in linear amplification, are calibrated on the previously trained weights and are controlled by a multi-current controller (Thorlabs, MLC8200CG), with average currents at 65 mA. The weighted signal is then multiplexed via the arrayed waveguide grating and fed to the optical nonlinear function, a nonlinear (NL) SOA-based wavelength converter (SOA-WC) with current at 120 mA, which converts the summation of the weighted inputs to another channel at 1546.72 nm, by exploiting the XGM. In order to emulate the evolution of the optical signal propagation from layer to layer, a noise source is coupled to the input signal to tune the input noise floor.

The converted signal is assessed after filtering it with an optical tunable filter with 1 nm passband width (Santec, OTF-950) and detecting it via an avalanche photodiode (APD) with -20 dBm sensitivity (Fiber-Photonics, APD-M-10-SMA-FA). The time traces are recorded on a digital phosphor oscilloscope (DPO, Tektronix, DSA-72004C). The performance of the neuron is evaluated by calculating the normalized root mean square error (NRMSE) between the measured output time traces and the expected time traces calculated starting from the utilized input time traces and the trained weighting factors (see in equation A1).

By tuning the noise source at the input, here we used an Erbium-Doped Fibre Amplifier (EDFA, PriTel Inc., LNHPFA-22), tuning the driving current of the booster EDFA, we can obtain the error evolution at the converted output, with respect to the OSNR at the input. Although the in-band noise of the converted output is not detectable at the measured output spectrum, the NRMSE on the detected electrical signal can be utilized as an indication of the OSNR at the output, as long as the response of the APD is determined with a back-to-back (B2B) measurement first to get the OSNR of the APD itself. By combining the error evolution measured with respect to the input OSNR together with the inverse relation of the NRMSE versus the OSNR, using the B2B measurement of the OSNR at the APD, we can determine the equivalent OSNR at the neuron output.

4.2.1. SOA characterisation

To determine the total ASE $\rho_{ASE,c}$ after WC-SOA in the converted output signal, using equations (4.1) and (4.2), we need to obtain the conversion efficiency η_i, the input SSE $\rho_{sse,i}$ and output ASE $\rho_{ASE,i}$. Therefore, we need to measure the gain response of the pre-amplifier SOA, weighting SOA and WC-SOA. The gain and ASE for WC-SOA and pre-amplifier SOA are easily assessable because they are at the input and output of the chip. Since the pre-amplifier SOA and the weighting SOAs are identical, within the same PIC and both with 1 mm length, the response of the weighting SOAs will be considered to be identical to the response of the input pre-amplifier SOA, which is assessable in the experiments.

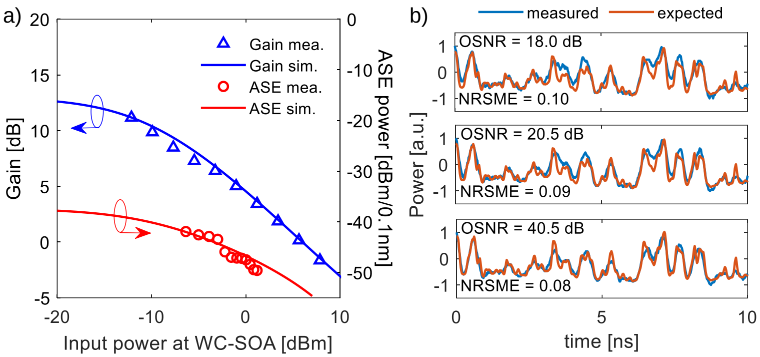


Fig. 4.2 (a) The gain (blue) and ASE power (red) vs. input power at WC-SOA, from experiments (scatters) and simulation (solid line) .(b) The times measured (blue lines) at the output of all-optical neuron with 7 channel input, and the input OSNR setting at 18.0, 20.5, and 40.5 dB, com-pared to the expected (red lines) time traces calculated from weighed addition.

To correctly represent the properties of the SOAs on chip, we use the on-chip SOAs as photo-detectors to measure the on-chip power after the pre-amplifier SOA and WC-SOA, entering from the input and output side, respectively, to determine the gain profile. Fig. 4.2 illustrates the gain and ASE as a function of the input power at the WC-SOA. The blue triangles present the measured gain data and the solid blue line plots the simulated response with unsaturated gain $G_0 = 13$ dB, saturation power $P_{sat} = 6$ dBm, and internal loss $\alpha' = 0.5$, using equation (4.B4) in Appendix 4.B. The red circles present the output noise power measured at the output of WC-SOA, while the red solid line plots the simulation result using inversion parameters n_{sp} = 7.3 and b_{sp} = 1 in equation (4.8). The variation of the noise at the output of WC-SOA is attributed to some reflections happening in the photonic circuit. Using the same method, the gain response of the pre-amplifier SOA is modelled with unsaturated gain $G_{0,pre\text{-}SOA}$ = 9.5 dB, saturation power $P_{sat,pre\text{-}SOA}$ = 7.8 dBm, normalized waveguide loss $\alpha'_{pre\text{-}SOA}$ = 0.6, $n_{sp,pre\text{-}SOA}$ = 3.6 and b_{sp} = 1. The obtained parameters are now used for the NRMSE estimation. The rest of the other parameters used in the simulations is listed in TABLE 4.1.

4.2.2. The error evolution of AON

In this section we want to analyse the error evolution for a single neuron and validate this via experimental results. For this reason, the time traces of the converted signal are recorded from the output of the PIC after 1 nm optical bandpass filter. These are then compared to the expected time trace calculated with the input reference signals multiplied by the calibrated weighting factors. In Fig. 4.2

TABLE 4.1 Simulation Parameters

symbol	description	value	Unit
G_{Rx}	Electrical gain at receiver	15.3	dB
M_{apd}	Multiplication factor, APD	10	
R_{apd}	Responsivity, APD	0.7	
F_e	Noise figure, Electrical amplifier	15	dB
B_e	Electrical bandwidth, APD	10	GHz
B_o	Optical Bandwidth	125	GHz
P_{s0}	B2B reference input power	-15	dBm
N_{th}	Receiver thermal noise	4.45×10^{-8}	A^2
r	Input optical extinction ratio	15	dB
G_0	unsaturated gain, WC-SOA	13	dB
$G_{0,\ pre\text{-}SOA}$	unsaturated gain, pre/w-SOA	9.5	dB
P_{sat}	saturation power, WC-SOA	6	dBm
$P_{sat,pre/w\text{-}}$	saturation power, pre/w-SOA	7.8	dBm
α'	normalized waveguide loss, WC-SOA	0.5	
$\alpha'_{pre/w\text{-}SOA}$	normalized waveguide loss, pre/w-SOA	0.6	
$n_{sp,}$	noise inversion parameter, WC-SOA	7.3	
$n_{sp,pre/w\text{-}}$	noise inversion parameter, pre/w-SOA	3.6	
b_{sp}	noise saturation inversion parameter	1	
ω	Small signal modulation frequency	10	GHz
τ	Spontaneous carrier lifetime	200	ps
$a_{1,\ pre/w\text{-}}$	Quadratic coefficient, pre/w-SOA	-2.3×10^{-2}	nm^{-2}
$a_{1,\ wc\text{-}SOA}$	Quadratic coefficient, WC-SOA	-1.8×10^{-2}	nm^{-2}
$a_{2,\ pre/w\text{-}}$	Linear coefficient, pre/w-SOA	-7.3×10^{-3}	nm^{-1}
$a_{2,\ wc\text{-}SOA}$	Linear coefficient, WC-SOA	-1.6×10^{-4}	nm^{-1}
$\lambda_{c,pre/w\text{-}SOA}$	Centre wavelength, pre-SOA	1540.42	nm
$\lambda_{c,wc\text{-}SOA}$	Centre wavelength, WC-SOA	1548.70	nm
L_c	Coupling loss per facet	-1.5	dB
L_1	Loss from pre-SOA to w-SOA	-5.2	dB
L_2	Loss from w-SOA to WC-SOA	-8	dB
L_{OBPF}	Loss from optical bandpass filter	-7	dB
P_{cw}	Continuous wave laser power at WC-SOA	-13	dBm
$\rho_{sse,c}$	Noise power density of CW-laser at WC-SOA	-45	dBm /0.1nm
P_0	Input signal power per channel in experi-	-17.5	dBm

b, the blue lines depict the recorded time traces at the output of the 7-channel AON after filtering. These are superimposed to the expected output time traces shown in red lines. The calculated NRMSE from 7-channel AON is also indicated when setting the input OSNR at 18.0, 20.5 and 40.5 dB, with current of booster EDFA at 300, 120, and 30 mA. The OSNR setting is not smaller than 18 dB because the recent EDFA is saturated when driving current is higher than 300 mA.

In Fig. 4.3*a*, the B2B measurement at the photodetector is shown in black: The black solid line with crosses illustrates the NRMSE as a function of the input OSNR of the signal at the APD, measured by directly coupling the modulated signal to the receiver and setting the average optical input power at -15 dBm, since the output of the AON is ≥ -15 dBm after the 1 nm OBPF, in the experimental conditions. A lower input power will lead to higher error for the analogue optical input. The crossings show the measurement results, and the solid line plots the fitting result, obtained using equation (4.2). The errors in the flat region are due to the signal-spontaneous beating noise and the thermal noise of the receiver, while the increasing of error obtained when OSNR <30 dB is attributable to the spontaneous-spontaneous beating noise at the detection. The optimal agreement between the measurement and the modelled curve suggest that we have an accurate relation between the NRSME and the OSNR for the input signal at the photodetector. Later the fitted curve is used for further interpretation of the error evolution in the experiments. Obtained the B2B measurement at the APD and extracted the parameters needed for calculating the total ASE at the converted neuron output signal (from section 4.2.1), we can now simulate the error evolution of the AON. The blue, red, green, magenta scatters and solid lines in Fig. 4.3*a*

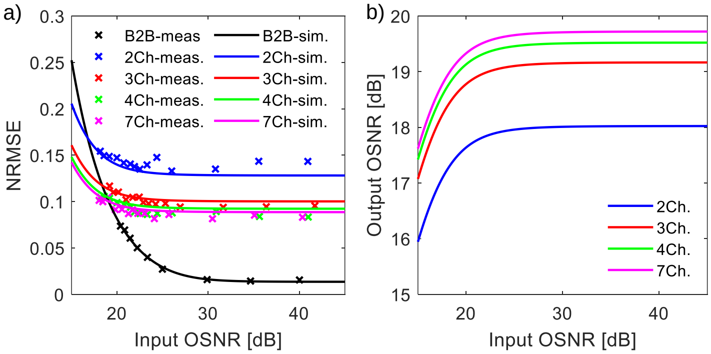


Fig. 4.3 (a) The measured (crosses) and simulated (solid line) NRMSE from the detection, for the back-to-back measurement at the photodetector (black), and all-optical neuron with 2, 3, 4, and 7-channel inputs. (b) The equivalent output OSNR (blue) and the noise figure (red) when tuning the input OSNR for 2, 3, 4, and 7-channel AON.

illustrate the measured and simulated errors when tuning the OSNR at the neuron input from 15 dB to 45 dB, for an AON with 2-, 3-, 4- and 7-channel inputs. The 5- and 6-channel curves are not shown as they are very close to the error trend for 7-channel input AON. When the input OSNR is >25 dB, the error stays almost flat, mainly due to the ASE noise coming from the WC-SOA, while a relatively small contribution comes from the noise at the neuron input. When the input OSNR is <25 dB, we can see a noticeable increase of the error, resulting in poorer performance of the AON, attributing to the conversion of input signal noise to the output channel on top of the ASE from the WC-SOA.

The data points shown in Fig. 4.3*a* for all the different number of inputs channels to the neuron are calculated as a function of the input OSNR using the measured time traces as shown in Fig. 4.3*a*. The experimental data points show a good agreement with the simulations, which we will use for representing the network error evolution. When comparing the error trend between the error evolution of the AON and the error from the B2B measurement at the APD, one can see that the error slope after the AON is flatter than the one of the B2B measurement. The increase of the error for the AON is slower than the original signal detection when decreasing the input signal OSNR. Specifically, when the input OSNR is smaller than the crossing point between the B2B error curve and the AON error curve, the errors at the AON output are smaller than the input errors. This suggest that the conversion of the input noises to the output signal is a fraction of the total noise present at the AON input, which means that the AON carries out noise compression. Seen from equation (4.17), there will be a OSNR that leads to $OSNR_{out} \geq OSNR_{in}$, i.e. an enhancement of OSNR when the input SSE noise becomes significant. This can also be understood by looking at the noise conversion from the input to the CW-Laser channel as shown in equation (4.3), (4.4): The conversion of the individual input channels to the output is limited to a fraction of the total conversion efficiency η_i, with both the signal and the noise converted to output. On the other hand, the input CW-channel is amplified by the WC-SOA with gain *G*.

To obtain the output OSNR, we used the B2B measurements as reference to determine the equivalent OSNR present on the receiver. The curves in Fig. 4.3*b* illustrates the output OSNR when tuning the input OSNR from 15 to 45 dB for 2, 3, 4, and 7-channel AON input, with blue, red, green, magenta colour. The output OSNR values are obtained by calculating the OSNR using the reverse function of B2B measurement in equation (4.2) with given NRMSE from the data for corresponding input OSNR, as shown in Fig. 4.3*a*. Using the same method, we can determine the output OSNR for all the input channels. The noise suppression effect is happened when input OSNR is smaller than OSNR around 19 dB, which gives a greater OSNR at the output than the input. This noise suppression provided by the SOA-based AON is beneficial to the scaling of the AON in the depth. Now we can emulate the scaling of the neural network in feedforward layer connections. The NRSME at the AON output has been obtained as shown in Fig. 4.3*a* as a function of the input OSNR at the first layer. Given an input OSNR, we then get the output OSNR as shown in Fig. 4.3*b*, which is then used as input OSNR to the second cascaded neuron, i.e. to the second layer. Using again the input OSNR we obtain the output OSNR, which is then used as input OSNR to the third cascaded neuron, i.e. the third layer, and so on and so forth. We can also emulate the number of neurons per layer by increasing the input channels to the neuron. With an M input channel, the performance of the AON is emulating a M neurons per layer to one neuron at the next layer.

4.2.3. The scalability of AON

In Fig. 4.3*a* we have presented the error evolution, with the B2B measurement as reference. To map the obtained error into a layer number, the changes in OSNR will be determined as plot in Fig. 4.3*b*. In Fig. 4.4*a*, the solid lines and dashed lines plot the error and the OSNR evolutions, respectively, when cascading neural layers with 4, 5, 6 and 7 neurons per layer (which is the same as saying with AON with 4-, 5-, 6- and 7-channel inputs) with a fixed input power of -17.5 dBm/channel. The error levels, after quickly increasing for a few numbers of consecutive layers, keep increasing but with a much smaller rate of change (slope). Similarly, the OSNR will converge to a minimum level. These maximum error levels and

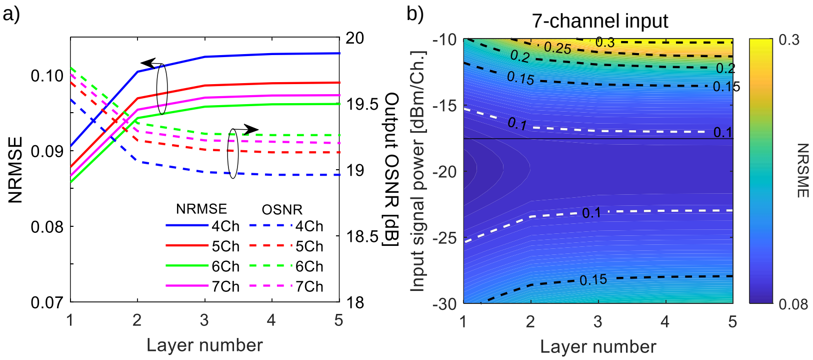


Fig. 4.4 (a) The NRSME (solid lines) and output OSNR (dashed lines) at the output of all-optical neuron with 4, 5, 6, 7 channel inputs as a function of the layer number. (b) The NRSME changes when tuning the input signal power from -30 to -10 dBm per channel, and changing the layer number from 1 to 5 layers, for an AONNs with 7-channel inputs. The black line presents the magenta line in (a).

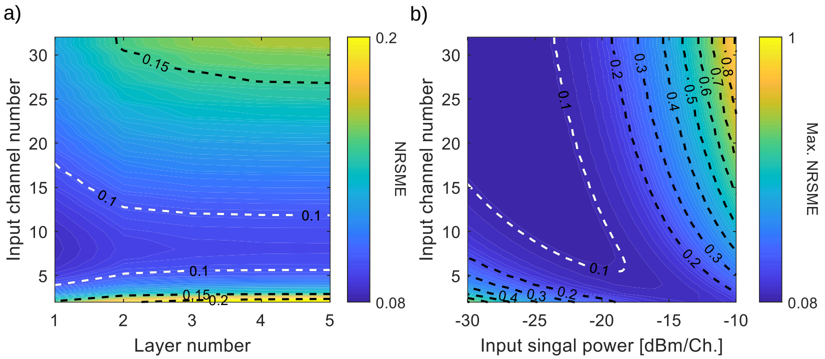


Fig. 4.5 (a) The NRSME changes vs. input channel numbers and layer numbers for fixed input signal power at -20 dBm per channel. (b) The maximum NRSME vs. the input channel numbers when tuning the input signal power from -30 to -10 dBm per channel.

minimum noise levels are defined by the crossing point shown in Fig. 4.3*a*: When the input OSNR is at the crossing point, the equivalent output OSNR will be the same as the input. Converting the WDM input into a single channel at the AON output via XGM, the influence of the input noise is compressed, resulting in a small rate of change of OSNR at the output of layer when increasing the number of cascaded layers. This effect is similarly visible for the error level or an arbitrary number of layer connection. On the other hand, one can note that the minimum output error is defined by the performance of the first layer: this depends on the ASE noise generated by the WC-SOA (equation (4.2)). The error decrease then increased when increasing the input channel number from 4- to 7-channels, which is attributing to the compromise of conversion efficiency and the gain at WC-SOA, known from equation (4.C3), the conversion efficiency is increased for increasing the input channel number while the gain is decreased when increasing the input signal power to the WC-SOA, seen from Fig. 4.2.

The input power to the WC-SOA is studied for the 7-channel inputs as a function of the layer numbers, to understand what the influence of the input power at the AON is on the resulting error when increasing the number of layers. In particular, the 3D mapping in Fig. 4.4*b* shows the NRMSE evolution with tuning the input signal power from -30 to -10 dBm and for layer scaling from 1 to 5. The black line crossing the map represents the input power used for 7-channels input AON in Fig. 4.4*a*. The contour lines in white identify the optimal input power region, around -20 dBm/channel, with an input dynamic range of about 6 dB for arbitrary layer scaling, for a maximum error of 0.1.

Furthermore, the error of the AON output when tuning the input channel number is investigated, to understand its influence on the network scalability.

Specifically, considering an input channel number larger than 7 channels with 400 GHz channel spacing (as used in the experiments), we have simulated the scaling with 100 GHz channel spacing. This means that for a 3-dB gain bandwidth of 32 nm, as measured from the optical spectrum at the WC-SOA output, the SOA is capable to amplify inputs of up to 32 channels. The equalisation of the input signal can be realized by slightly adjusting the input signal power or the current on the weighting SOA (before training). Fig. 4.5*a* depicts the NRSME error against input channel number and for layer scaling from 1 to 5, with fixed power of -20 dBm per channel. Here we see the errors will be above 0.1 when increasing the number of input channels. The contour for NRMSE $\leq$ 0.1 shows the operation regime of the AON, which suggest that the optimum number of input channels is 8, and that the AON can scale up to 18 channels for a single layer implementation, and up to 12 channels for an infinite layer connection, or in other words, a 12-input/neuron 12-neuron/layer neural network is feasible to be infinitely cascaded with expected NRMSE < 0.1.

To further improve the performance of the AONN for the SOA-based AON, the input signal power can be optimized respecting to each input channel number and enable scalability in the height of the network. Fig. 4.5*b* illustrates the maximum NRMSE of the AON for arbitrary layer number connection, obtained from the crossing point from Fig. 4.3*a*, with tuning input channel number from 2 to 32 channels and tuning the input signal power from -30 to -10 dBm per channel. The result shows that an optimum input signal power can be determined for individual input channel numbers, and the input dynamic range for NRMSE <0.1 can be greater than 10 dB when the input channel number is greater than 16 channels. Although using a higher number of input channels can improve the conversion efficiency, the optimized input signal power shifts to a lower value.

The increase of the input channel number may be limited by the bandwidth of the SOA as well as the performance of the AWG, which may induce extra loss and crosstalk when increasing the channel number. The reduction of input signal power will eventually be limited by the sensitivity of the weight amplifier, which defines the lowest input power. This may be solved by increasing the gain of the pre-SOA while reducing the gain of weight SOA to keep the input higher than the sensitivity of the weight SOA, albeit match to the total input power to the WC-SOA. Moreover, apart from the optimisation of input power, the AON output error can be further reduced by improving the noise inversion parameter n_{sp} of WC-SOA, since the output ASE from the WC-SOA defines the lower bound of the NRMSE.

In Appendix 4.E, the AONN operation with multi-level input is demonstrated to utilize the analogue nature of the light for optical signal processing to provide even higher throughput improvement.

4.3. Conclusion

The scaling of all-optical neural network is experimentally demonstrated with photonic integrated SOA-based AON, utilizing XGM in an SOA as nonlinear transfer function. A noise model is developed for simulating the noise accumulation of the AON. The model shows good agreement with the experimental data, and it actually relies on the characteristic parameters of the used SOA components. The data are analysed to interpret the scalability of the AON in terms of input channel number and layer number. The results show that the WDM input, entering the non-linear function and realising *N:1* conversion on a single output, undertakes noise compression, which defines a maximum error at the output of the AON. And the recent AON structure is capable to establish a 12-input/neuron 12-neuron/layer arbitrary layer number all-optical neural network, with a finial NRMSE < 0.1, with optimized input signal power at -20 dBm per channel, for a channel spacing of 100 GHz and a gain bandwidth of 32 nm. The noise model can be further used to investigate other parameters for the *N:1* XGM-based conversion in optical signal processing, like noise inversion parameter, passive losses, and SOA gain response, etc. In conclusion, utilizing WDM-input-to-single-output conversion via XGM in SOA, the proposed AON structure can possibly scale up to an arbitrary layer number connection with large input channel number, resulting in acceptable maximum output signal error.

Appendix 4

4.A. Back-to-Back measurement

In this section, the format of the NRSME vs. input noise with fixed input power is derived. The definition of the NRSME in this paper is:

$$NRSME = \frac{\sqrt{\sum_{i=1}^{m}(x_i - E_i)^2/m}}{I_{s,max} - I_{s,min}}, \tag{4A.1}$$

where x_i is the *i*-th measured data value and E_i is the *i*-th expected value for input reference with calibrated weighted addition. m is the number of recorded data points. $I_{s,\,max}$ and $I_{s,\,min}$ are the maximum and minimum photocurrent on the detection. Considering the noise at the photodetector has a Gaussian distribution $\mathcal{N}(0, N_e)$, with N_e as the electrical noise at the receiver, equation (4A.1) becomes:

$$NRSME = \sqrt{N_e}/\left(I_{s,max} - I_{s,min}\right). \tag{4A.2}$$

Equation (4.1) is obtain with substituting $S = (I_{s,max} - I_{s,min})$ as the span of the detected photocurrent.

Since the B2B error curve is measured directly tuning the noise at the signal for the receiver, we need to know the noise from the detection. At the receiver, the noises are [101]:

$$N_{shot} = 2eB_e\left(I_s + I_{sp}\right), \tag{4A.3}$$

$$N_{s-sp} = 4I_sI_{sp}B_e/B_o\,, \tag{4A.4}$$

$$N_{sp-sp} = I_{sp}^2B_e(2B_o - B_e)/B_o^2, \tag{4A.5}$$

$$N_{th} = I_{th}^2, \tag{4A.6}$$

$$N = N_{shot} + N_{s-sp} + N_{sp-sp} + N_{th}, \tag{4A.7}$$

where the noise N consists of shot noise N_{shot}, signal-spontaneous beating noise $N_{s\text{-}sp}$, spontaneous-spontaneous beating noise $N_{sp\text{-}sp}$, and thermal noise N_{th}. e the electric charge, B_e is the electrical bandwidth, B_o is the optical bandwidth, I_s and I_{sp} are the photocurrent from the signal and noise at the receiver:

$$I_{s,max/min} = R_0P_{s,max/min}, I_{sp} = R_0P_{sp}F_e, \tag{4A.8}$$

with R_0 is the responsivity of the photodiode. In case of APD and followed with an electrical amplifier, $R_0 = M_{apd}R_{apd}G_{Rx}$, where M_{apd}, R_{apd} are the multiplication factor and responsivity of the APD when $M_{apd} = 1$ and G_{Rx} is the electrical gain in the receiver circuit. And F_e is the noise figure of the electrical amplifier. I_{th} is the thermal current. B_e is the electrical bandwidth, and B_o is the optical bandwidth.

For an extinction ratio r of received signal, set I_s as averaged photocurrent:

$$I_{s,max} = rI_{s,min}, \tag{4A.9}$$

$$I_s = \left(I_{s,max} + I_{s,min}\right)/2\,, \tag{4A.10}$$

yields,

$$P_s = \ (P_{s,max} + P_{s,min})/2\,, \tag{4A.11}$$

$$S = 2I_s(r-1)/(r+1). \tag{4A.12}$$

From equation (4A.2),

$$NRSME = \sqrt{N_e}/S\,. \tag{4A.13}$$

Substituting (4A.8) and (4A.11), with fixed r and I_s, equation (4.2) is obtained:

$$NRSME = \ \sqrt{c_1 I_{sp}^2 + c_2 I_{sp} + c_3},$$

with,

$$c_1 = \ B_e(2B_o - B_e)/(B_o^2 S^2), \tag{4A.14}$$

$$c_2 = 2(eBe + 2I_s Be/Bo)/S^2\,, \tag{4A.15}$$

$$c_3 = \left(2eBeI_s + I_{th}^2\right)/S^2\,. \tag{4A.16}$$

4.B. Conversion Efficiency

Here calculation of conversion efficiency at the WC-SOA is calculated, from WDM input to single channel output using the small signal modulation method. Considering WDM inputs with small signal modulation p_i modulated on average power P_i for the *i*-th input signal at WC-SOA. For cross gain modulation, the conversion from *j*-th input channel to *i*-th output channel after the WC-SOA (with length of L). The conversion efficiency from *z=0* to *z=L*, i.e. from $p_i(0)$ to $p_k(L)$, along the length of WC-SOA, can be calculated as [98]:

$$\eta_{ki} = \left|\frac{p_k(L)}{P_k(L)} \Big/ \frac{p_i(0)}{P_i(0)}\right|$$

$$= \left|\frac{p_k(0)P_i(0)}{P_k(0)p_i(0)} - \frac{p_T(0)P_i(0)}{P_T(0)p_i(0)}F(L)\right|, \qquad (4B.1)$$

with $p_T = \sum p_i$ and $P_T = \sum P_i$.

And

$$F(L) = 1 - e^{-K(L)}, \qquad (4B.2)$$

$$K(L) = \frac{1}{1 - j\omega\tau\alpha'}\left\{\alpha' \ln\frac{G_0}{G} - \ln\left[1 - \frac{(G-1)P_T(0)/P_{sat}}{1 + GP_T(0)/P_{sat} + j\omega\tau}\right]\right\}, \qquad (4B.3)$$

with internal loss $\alpha' = \alpha/\Gamma g_0$, the normalised waveguide loss coefficient, P_{sat} is the saturation power, ω is the small signal modulation frequency, τ is the carrier lifetime, and j denotes the imaginary unit. And the unsaturated gain:

$$G_0 = \exp[(\Gamma g_0 - \alpha)L].$$

The amplifier saturated gain is defined from [73][98]:

$$\alpha' \ln\frac{G_0}{G} = \ln\left\{\frac{1 - \alpha'[1 + P_T(0)/P_{sat}]}{1 - \alpha'[1 + GP_T(0)/P_{sat}]}\right\}. \qquad (4B.4)$$

Assuming the modulation index r' of the input WDM channels are the same:

$$r' = \frac{p_i(0)}{P_i(0)}, \text{for } i = 1, \ldots, M. \qquad (4B.5)$$

With $p_i(0) = P_{s,max} - P_{s,min}$, and $P_i(0) = P_s$, in equation (4.A9), the modulation index *r'* is related to the extinction ratio *r*:

$$r' = 2(r-1)/(r+1) \qquad (4B.6)$$

And the percentage of the WDM channel is defined by the normalized weights on the linear unit,

$$w_i' = \frac{p_i(0)}{p_T(0)} \qquad (4B.7)$$

And the small signal modulation at the CW laser input is $p_c(0) = 0$, substituted in (B1), the conversion efficiency from i-th input to the converted channel is:

$$\begin{aligned}\eta_{ci} = &\left|\frac{p_c(L)}{P_c(L)}/\frac{p_i(0)}{P_i(0)}\right| \\ = &\left|\frac{p_T(0)P_i(0)}{P_T(0)p_i(0)}F(L)\right| \\ = &\left|\frac{P_i(0)}{P_T(0){w'}_i}F(L)\right|. \qquad (4B.8)\end{aligned}$$

Denoting $\eta_i = \eta_{ci}$ in equation (4.B4) and define $P_i(0)$ at the input as P_i, we obtain equation (4.2) in the main text.

4.C. NRMSE versus Input Noise

Here the format of the output NRSME vs. input noise, is determined, with amplified converted signal and changed extinction ratio.

The modulation index of the output after conversion, for the amount from the i-th input channel, from equation (4.B1), is:

$$\frac{p_{c,i}(L)}{P_c(L)} = \eta_i \frac{p_i(0)}{P_i(0)}. \qquad (4C.1)$$

With (B5), the modulation index r_c' at the converted output is:

$$\begin{aligned}r_c' = \frac{p_c(L)}{P_c(L)} &= \sum_{i=1}^{M} \frac{p_{c,i}(L)}{P_c(L)} \\ &= r' \textstyle\sum_{i=1}^{M} \eta_i . \qquad (4C.2)\end{aligned}$$

So that the total conversion efficiency is:

$$\eta = r_c'/r' = \textstyle\sum_{i=1}^{M} \eta_i . \qquad (4C.3)$$

Consider the referenced signal span in the B2B measurement is $S_0 = 2R_0P_{s0}r'$, with averaged reference optical power P_{s0}, the span of the output channel S_c is:

$$\begin{aligned}S_c = 2R_0P_c(L)r_c' &= S_0P_c(L)r_c'/(P_{s0}r') \\ &= \eta S_0 P_c(L)/P_{s0} = \eta S_0 G'. \qquad (4C.4)\end{aligned}$$

If $G' = P_c(L)/P_{s0} = 1$, i.e., the same average power at the output, the *NRSME* will be $1/\eta$ times the original value. If $G' \neq 1$, there is slightly gain/loss at the output compared to the reference power. Both the ASE and signal power will be G' times the original noise and signal:

$$I_{sp,c} = G' I_{sp} \tag{4C.5}$$

$$\begin{aligned} I_{s,c} &= R_0 P_c(L) = R_0 G' P_{s0} \\ &= G' I_s \end{aligned} \tag{4C.6}$$

substitute (4C.4) - (4C.6) in equation (4.2) with (4A.14) - (4A.16) the $NRSME$ will be:

$$NRSME = \sqrt{c_1' I_{sp,c}^2 + c_2' I_{sp,c} + c_3'}/\eta\,,$$

With

$$c_1' = B_e(2B_o - B_e)/(B_o^2 S_0^2), \tag{4C.7}$$

$$c_2' = 2(Be/G' + 2I_s Be/Bo)/S_0^2\,, \tag{4C.8}$$

$$c_3 = \left(2BeG' I_s + I_{th}^2\right)/(G' S_0)^2\,. \tag{4C.9}$$

Substituting $I_{sp,c}$ with $I_{wc\text{-}SOA}$, we can obtain equation (4.10).

4.D. Input Noise Suppression Condition

Here we derive the condition to achieve the noise suppression, which defines the input OSNR at WC-SOA when achieving an enhanced output OSNR.

From equation (4.13)-(4.15), we have,

$$\frac{GP_{cw}}{\rho_{ASE} B_o} \geq \frac{P_{in}}{B_o \rho_{in}} = \frac{\sum P_i}{B_o \sum \rho_{sse,i}}, \tag{4D.1}$$

From equation (4.3), unfold the CW channel, it yields,

$$\begin{aligned} &\frac{GP_{cw}}{\sum \eta_i \rho_{sse,i} + \sum \eta_i \rho_{wc-ASE,i}/G + G\,\rho_{sse,cw} + \rho_{ASE,cw}} \\ &\geq \frac{\bar{P}_{in}}{\bar{\rho}_{sse}}. \end{aligned} \tag{4D.2}$$

With $\bar{P}_{in} = \frac{1}{M}\sum P_i$, $\bar{\rho}_{sse} = \frac{1}{M}\sum \rho_{sse,i}$, consider in a simple case, $\eta_i = \eta/M,$ i.e. the conversion efficiency is the same for all the input channel to WC-SOA, for all input signal power is identical as $\bar{P}_{in}$. With $\bar{\rho}_{ASE} = \frac{1}{M}\sum \rho_{ASE,i}$, after some algebra, we obtain,

$$\bar{\rho}_{sse} \geq \frac{\eta\bar{\rho}_{ASE}/G + G\,\rho_{sse,cw} + \rho_{ASE,cw}}{\frac{GP_{cw}}{\bar{P}_{in}} - \eta} \qquad (4D.3)$$

Therefore,

$$\frac{\bar{P}_{in}}{\bar{\rho}_{sse}B_o} \leq \frac{GP_{cw} - \eta\bar{P}_{in}}{\eta\bar{\rho}_{ASE}/G + G\,\rho_{sse,cw} + \rho_{ASE,cw}} \cdot \frac{1}{B_o} \qquad (4D.4)$$

With $OSNR_{in} = \frac{\bar{P}_{in}}{\bar{\rho}_{sse}B_o}$, the equation (4.17) is obtained. And seen from equation (4D.3), equation (4D.4) exits only if $\frac{GP_{cw}}{\bar{P}_{in}} \geq \eta$.

4.E. All-optical Neuron Operation with Multi-level Inputs

In this study, we implement an SOA-based neuron with off-the-shelf components to study the performance of the linear part and nonlinear part of the all-optical neuron with respect to multi-level amplitude modulated inputs and with NL-SOA based optical nonlinear transfer function (NL-TF). The analogue linear and non-linear computation accuracy is studied with respect to the binary data computation from the computer in order to identify data format regimes with optimal neuron accuracy.

Fig. 4E.1*a* shows the experimental setup. A WDM laser source is operated with four wavelengths at ITU channel 21, 23, 25, and 27. These are multiplexed to be modulated by a PAM modulation format up to 512 levels on a single modulator, driven by an arbitrary waveform generator, with the multi-levels converted from pseudo random bit sequence (PRBS). After decorrelated using varied lengths of optical fibers, the WDM signal is fed into the neuron, with a power of -10 dBm per channel. Fig. 4E.1b presents the schematic of the all-optical neuron. In the blue line box, the linear weighted addition part (Σ) is realized by exploiting the linear gain region of weight-SOAs: the multiplications of all the input channels with fixed weights are filtered and summed up via an AWG before being sent to the optical NL-TF (red line box). The NL-TF (φ) is implemented via an SOA-based wavelength converter, with a CW laser at λ_{TL}=1554.13 nm, with a power of -3 dBm. This activation function receives the WDM input signal and converts it into λ_{TL}. The full neuron is programmed electronically with high-speed drivers which receive a control signal from an FPGA interfaced to a computer (Fig. 4E.1*a*). Control currents in weigh-SOAs are tuned up to 70 mA to compensate for the path loss and to set different weight factors to the input data. The NL-SOA current is set at 100 mA. The neuron is implemented via off-the-shelf optical

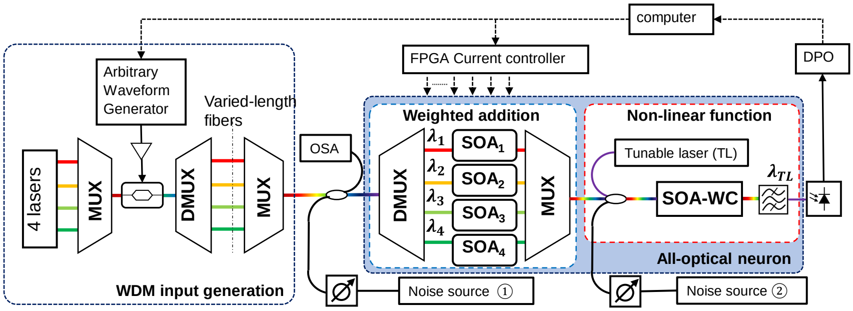


Fig. 4E.1. The experimental setup for scalability investigation of the SOA-based all-optical neuron with multi-level, multi-wavelength inputs.

components to access each probing point and understand linear and non-linear function outcomes as a function of the channel numbers and the modulation level numbers. The neuron output, detected by a linear photodetector, is recorded by a digital phosphorous oscilloscope (DPO) and post-processed. In the future, the neuron will be part of a full layer of n neurons with m input wavelengths where

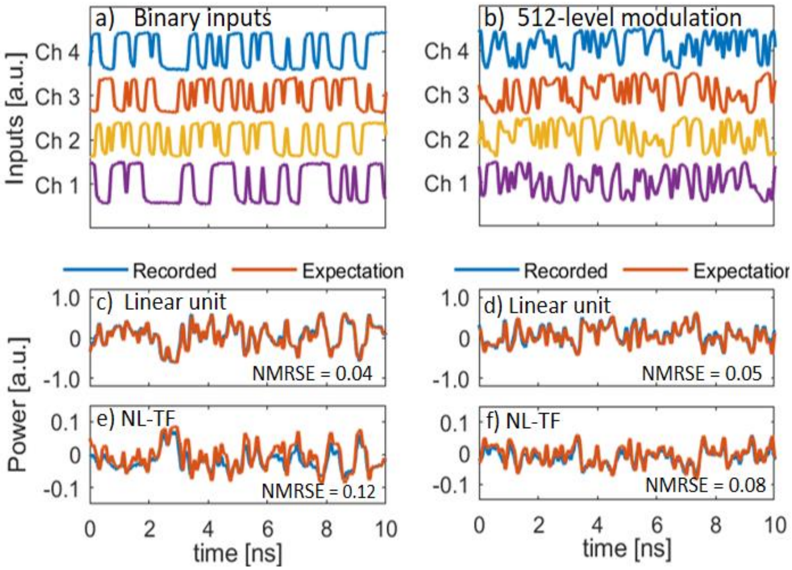


Fig. 4E.2. Four-channel addition for (a) binary signal and (b) 4-level amplitude modulation. (c, d) Output signal detected after the linear weighted addition unit and (e, f) the nonlinear function for the binary signal in (a) and the multilevel signal in (b), respectively.

both the layer input and output are WDM signals, by multiplexing the neuron outputs (See Fig. 3.1).

Four-channel inputs with multilevel modulation from 1 to 9 bit/symbol (from 2 up to 512 levels) are generated at a baud rate of 10 Gbaud/s. Fig. 4E.2 shows in particular the case of 4 binary channels and 4 multi-level input channels with 9 bit/symbol (Fig. 4E.2 *a* and *b*) at the input of the optical neuron, in 10 ns time window. Both the linear unit output and the non-linear output of the neuron are shown for the 2 cases, respectively (Fig. 4E.2*c,e* and Fig. 4E.2*b,d*), resulting in NRMSE of 0.04 and 0.12 for the binary signal and 0.05 and 0.08 for the multilevel signal, respectively. The recorded output is shown in blue lines and the expected outputs are plotted in red. It is evident how while the linear function linearly sums up the inputs, irrespective of the modulation format, the non-linear function acts on the input signal in a different manner: a multi-level signal helps the non-linear function to follow smoothly, reducing the errors. This is because higher multi-level signals provide more levels at the input of the NL-TF, so relaxing the demand for rapid changes in the carriers of the NL-SOA.

The impairment of the output signal after the NL-TF is coming from the non-linearity of the NL-SOA. A pre-recorded NL-TF can be exploited to calculate the expected output of the all-optical neuron. Fig. 4E.3*a* illustrates the non-linear function recorded by plotting the correlation mapping from the input to the output of the NL-SOA, with the blue crosses depicting the original data and the polynomial fitting showed via a red line. The transfer function shows an inverted relation between the input and the output of the NL-SOA with a reversed cube curve shape, which is attributable to the cross-gain modulation of the NL-SOA. Fig. 4E.3b shows the response of NL-SOA when changing input channel numbers from 1, 2 to 4, with lines in blue, red and yellow, respectively, when the inputs are modulated with binary signals. The transfer function tends to be linear when adding more input channels. This might be due to an increase of the signal amplitude levels at the output of the weighted addition unit, as well as an increase of the total power of the probe signal of the NL-SOA wavelength converter improving the conversion efficiency since the total probe power changes from -12 dBm for one channel to -9.5 dBm for four-channel weighted addition. To improve the non-linear response of the NL-SOA, the probe laser power should be tuned to drive the NL-SOA in the nonlinear regime.

After recording the NL-TF of the neuron, the performance of the neuron can be defined by the accuracy of the neuron output. We change the modulation levels from the 1 to 9 bit/symbol and record the output of the linear weighted addition unit, as well as the output of the NL-TF. Fig. 4E.4 illustrates the error evolution when increasing the modulation levels. Fig 4E.4 *a-c* plot the NRMSE variation

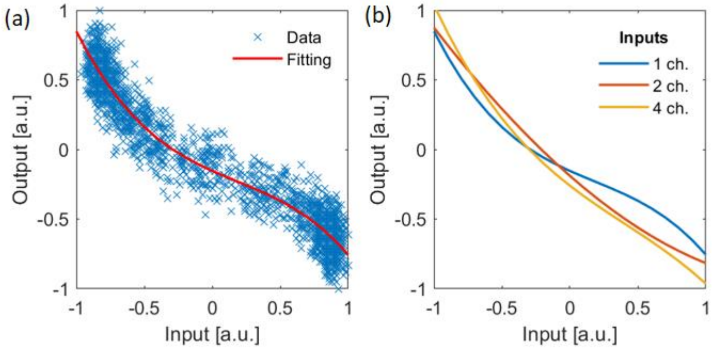


Fig. 4E.3. (a)The nonlinear function with the input-output correlation (blue crosses) for one channel input, and the polynomial fitting (red line). (b)The NL-TF curves with inputs of 1 channel (blue), 2 channels (red) and 4 channels (yellow).

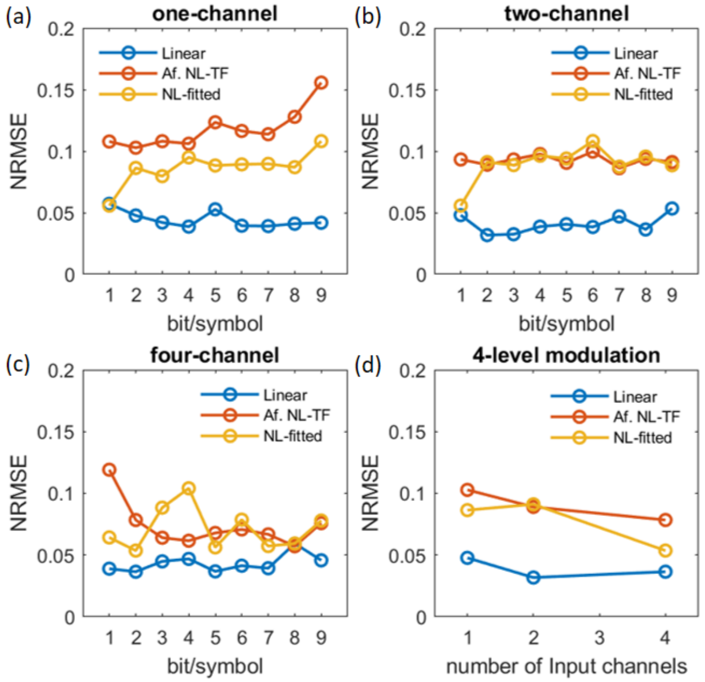


Fig. 4E.4: Error evolution when changing the multi-level modulation from 1 bit/symbol to 9 bit/symbol, for (a) one channel, (b) two channel, and (c) four channel weighted addition. (d) The error variation when increasing the channel number with 4-level (2 bit/symbol) modulation.

for 1, 2, and 4-channel input, respectively. The blue lines represent the NRMSE of the linear part only, the red lines show the NRMSE of the non-linear part only, and the yellow lines plot the NRMSE of the non-linear part, when considering the nonlinear response of the NL-SOA recorded in Fig. 4E.3.

The blue lines are almost flat, which suggests that SOA-based weighting is not notably influenced by the modulation levels. The stable performance of the linear weighted addition is also attributable to the high extinction ratio of the weight-SOA and the noise filtering at the AWG. The red and the blue lines instead show how the nonlinearity from the NL-SOA induces distortion on the linear signal. The errors after the NL-TF are flat for one-channel and two-channel after conversion but they decrease for four-channel addition when increasing the modulation level number. This suggests that the nonlinearity of the NL-SOA tends to relax when a higher number of levels is used in the modulation, which might due to the smoothness of the output signal variation. The nonlinearity may be improved by changing the operation regime of the NL-SOA. When looking at the yellow lines, for one channel and two channels the error is reduced comparing to the red line error, however for four-channel inputs, the nonlinearity is not correctly included in the expectation signals, which is attributable to the need of an ad-hoc NL-TF for different operating cases or to the need to optimize the nonlinearity of the NL-SOA itself. However, the average error after the nonlinear function is still 0.07 for this SOA-based all-optical neuron, which is comparable to the case with E/O conversion, as shown in section 2.2. In particular, for a 9 bit/symbol, the error after the NL-TF is 0.08. Finally, Fig. 4E.4*d* plots the error changes for 4-level input modulation when increasing the number of input channels into the neuron. The error is seen to slightly decrease when input channel numbers increase. This is due to a parallel increase of the input power at the linear part output/nonlinear input.

Chapter 5

Photonic Convolutional Neural Network

In this chapter, the convolutional neural network is introduced in section 5.1. The optical CNN in free space with a low cost 4-f system is demonstrated in section 5.2, which avoids micro-optics displacement constraints and trains on less data, while obtaining good accuracy. Then the implementation of photonic convolutional neural networks (PCNN) on-chip is discussed, in section 5.3, by exploiting the cross-connect architecture shown in Chapter 2, with the exploitation of both the wavelength division multiplexing and the time multiplexing schemes. Therefore, the performance is projected in terms of computational speed and energy efficiency in comparison with state-of-the-art demonstrated PCNNs on chip.

5.1. Convolutional Neural Network

The convolutional neural network (CNN) is one of the promising feed-forward neural networks for image classification, voice recognition and nature language processing. It is widely used in the research such as feature extraction, DNA-protein bonding prediction, etc. With the help of highly parallel computing, CNNs can process the input data faster, however they require great computing power for the matrix multiplications, just as the other deep neural networks as mentioned in Chapter 1. Off-the-shelf electronics such as GPUs and TPUs, performs 2D convolutions by running an iterative algorithm, consisting of point wise multiplication and kernel shifting. Fig. 5.1 shows, for ease of understanding, the convolution operation with a 4×4 2-dimensional image and a 2×2 kernel. Fig. 5.1*a* and *b* shows how the 1st and 2nd input 2×2 matrixes, red and green boxes of the input image, are multiplied by the elements of the kernel, shown in yellow, with element-wise multiplication, and then summed up to obtain the 1st and 2nd element of the output, shown in red and green at the output, respectively. One can see the computational complexity of this scanning computation is $O(n^2k^2)$ for the convolution operation of an $(n\times n)$ image and a $(k\times k)$ kernel, suggesting that computing

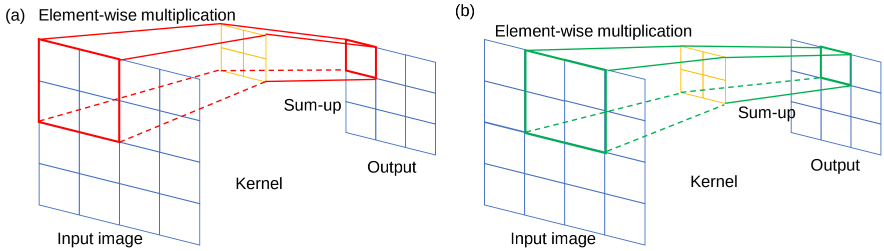


Fig. 5.1 Schematic diagram for convolution operation for (a) the first (red) and (b) the second (green) elements of the output.

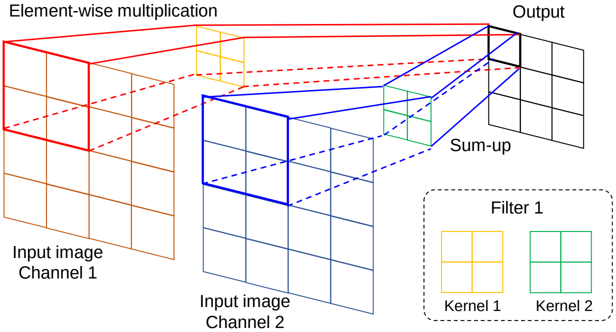


Fig. 5.2 Schematic diagram for convolution operation for multi-channel operation with a filter consisted of 2 kernels.

of the 2D convolutions scale quadratically with the size of the input images, resulting in higher power consumption and longer execution times.

Moreover, for the input image with multi-channels, e.g., an RGB colour image, containing images in 3 channels (RGB channels), the output of the convolution will be the summation of all the element-wise multiplication for different kernels of one filter. Fig. 5.2 sketches the first step of the convolution for 2-channel input image: channel 1 and channel 2 represent different data "dimensions". Moreover, there is one filter consisted of two kernels, shown in yellow and green colour. The first 2×2 matrix from the 1st channel image is multiplied by the kernel 1 and the first 2×2 matrix from the 2nd channel image is multiplied by the kernel 2. Then the first element of the output matrix is already the summation of all the multiplication and the convolution is carried out by scanning of input images as shown in Fig. 5.1. Specifically, each kernel of a filter catches certain features from the different channels. The output results to be still one matrix, but with lower dimension for a kernel scanning within the input image.

Photonic convolutional neural network is gaining attention, owing to the natural parallelism of light. PCNN based on free-space optics [102], fiber optics

[103],[104], and integrated optics [105]-[108] have been proposed to demonstrate the advance of convolutional operation utilizing optical beam parallelism in spatial, modal and frequency domains. Moreover, the first examples of PCNN on chip [106],[108] are also shown in great performance. The PCNN processor in [108], exploiting optical comb laser to enable dense WDM throughput, as a 4×4 kernel device with phase change materials embedded on cross-bar connections (Xbar-PCM), has shown computing speed 64 GMAC/s with 2 Gbit/s inputs.

In this chapter, an optical CNN system, with free-space optics performing Fourier transform, is demonstrated with accuracy comparable to state-of-the-art optical CNNs (section 5.2), and possible operations of the integrated cross-connect for convolutional operation are proposed and discussed (section 5.3). Specifically, in section 5.3 we introduce another dimension which is commonly implemented in electronics: The input channels and the multi-filters operation. The photonic integrated implementation so far have been mostly focusing on the single kernel operation [102]-[107], with the additional channel dimension implemented in the electrical domain [102]-[106]. Recently an integrated PCNN has been designed to process multiple-filters in the optical domain, demonstrating high throughput with WDM signals [108]. The PCNN is expected to be accelerated further by adding another dimension for parallel computing to the chip architecture. The cross-connect structure used in Chapter 2 is considered to be a good candidate to boost further parallel computation, as will be explained in section 5.3.

Note that in the translation of the multi-dimension, multi-kernel electronic concepts of a convolutional neural networks into a photonic integrated implementation we will call 'channel' (the additional dimension of virtual data used in the electronic calculation) as 'D_i' for the *i*-th channel input images (not to get confused with the optical input channels). Moreover the input optical channel will be referred as λ_i (to avoid confusion with the channel-dimension in electronics).

5.2. Optical CNN with a 4-f System

5.2.1. Experimental setup

Fig. 5.3*a* depict the conceptual schematic of the 4-f optical correlator system, which allows to perform the convolution operations by exploiting Fourier transform, with in total 4 times focal length of the used optical lens for the setup from input to output plane [109]. The expended (with lens L1 and L2) laser beam is used to generate the input images. After the first spatial light modulator (SLM, LM1), the Fourier transform of the input image is obtained by optical lens L3. The 2nd SLM is used in the Fourier plane to carry out matrix multiplication in

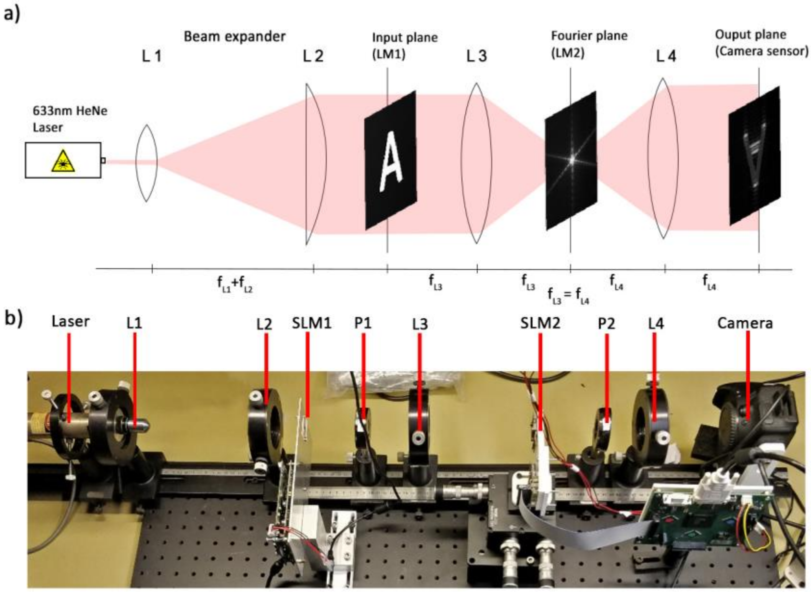


Fig. 5.3 (a) Schematic of the 4-f optical correlator system. (b) Experimental setup of the realized 4F optical correlator using off-the-shelf components. [109]

spectral domain. The optical lens L4 carry out the inverse Fourier transform, and the camera capture the output image at output plane. The convolution between an input image $f(x,y)$ and a kernel $k(x,y)$ in the spatial domain is equivalent to their product in the spectral domain:

$$h(x,y) = f(x,y) * k(x,y) = \mathcal{F}^{-1}\left(\mathcal{F}\big(f(x,y)\big) \cdot \mathcal{F}\big(k(x,y)\big)\right) \tag{5.1}$$

where * denotes the convolution operation, $\mathcal{F}$, the Fourier transform and $\mathcal{F}^{-1}$, the inverse Fourier transform. The inverse Fourier transform of modified spectral image returns the final convolutional results in spatial, therefore, the 2D spatial convolutions can be implemented with a 4-f correlator to perform one-step calculation optically. The direct usage of the kernels at the Fourier plane can reduce the training time of the CNN, which avoids the serial kernel scanning of the input image, as well as avoids the Fourier transform and inverse Fourier transform of the spatial kernels for the training. Hence, a 4-f system can reduce the computational complexity of any matrix convolution from O(n^2k^2) to O(1) [110], and reduce the training time. It means that the convolution operation will always consume a constant amount of computing time and power, for the possible sizes of the input and kernel matrix, which is defined by the size of the spatial light modulators used to implement the input and kernel.

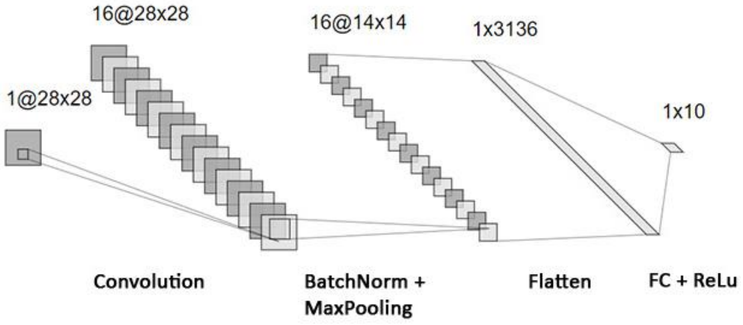


Fig. 5.4 Convolutional neural network structure implemented for MNIST classification.

Fig. 5.3*b* shows experimental setup of the optical CNN system. The 633nm He-Ne laser beam is expanded by a Keplerian beam expander (lenses L1 and L2) to cover the active area of the first spatial light modulator (SLM1) where an image is encoded will 2-D images from the computer. The optical lens L3 provides the Fourier transform at the focal distance as a Fourier plane. At the Fourier plane, 2nd SLM is utilized as a programmable filter where the Fourier transform of different kernels are displayed during inference. And polarizers P1 and P2 after the SLM are used to filter out the scattered light from the SLM and enhance the extinction ratio of the modulation of the optical beam. The optical lens L4 gives the inverse Fourier transform of the modified spectral on the output plane at the focal distance of L4, where a camera is employed to capture the output image and converted from the optical to the electronic domain for further processing on the computer.

To compare both accuracy and speed of a GPU-based CNN to the optical CNN, a simple CNN architecture, as shown in Fig. 5.4, including Convolution layer, Batch Normalization, Max Pooling, Flatten, and a 2-layer full-connected (FC) layer with ReLu function, is established to solve the MNIST classification problem [80]. The training of the CNN is carried out using the open-source machine learning library PyTorch [111], with a subset of roughly 300 images of the MNIST dataset in this demonstration. To train the CNN, a custom (electronic) Fourier convolution function was programmed which ensures that the kernels are initialized and trained in the Fourier domain [112], which can be direct used in the experiments and saves computation time without the kernel Fourier transformation. Testing showed that this custom function had similar accuracy results in a CNN as PyTorch's built-in 2D convolution function. The inference of the convolutional layer is demonstrated in the optical domain with the 4-f CNN system and benchmarked to the GPU-based convolutional computing.

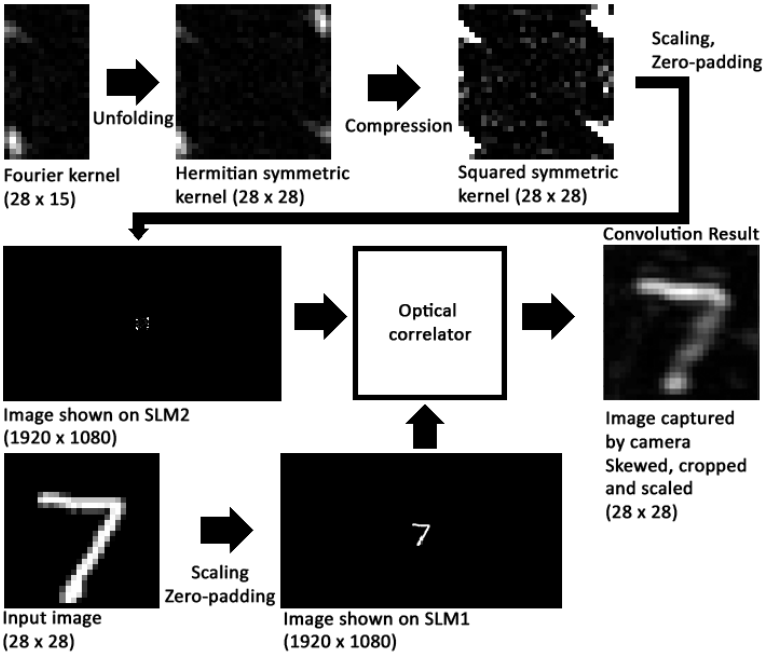


Fig. 5.5 Procedure to perform inference on the optical correlator, with experimental captured kernel/images [109].

Fig. 5.5shows the procedure of the operation of the 4-f CNN system. The spatial kernel is real-valued therefore the kernel in frequency domain is Hermitian symmetric, hence, the kernel is trained on the positive half and can be unfolded to obtain a square matrix, which can be displayed at the spectral filter on SLM2. A compression function is applied to enhance the light amplitude transmission through SLM2. Finally, both the input image and the squared symmetric kernel are scaled and zero-padded so that the images are displayed on the centre of the light modulators. And the camera detects the convolutional output image as shown in Fig. 5.4 and sends the convolutional output to the computer, which processes the data with the full-connected and provides the MNIST digits prediction.

5.2.2. Results

A comparison between the convolution results from electronic processor (using the custom Fourier convolution function) and the optical correlator system is shown in Fig. 5.6. Both convolution methods highlight the same areas of the picture, which indicates that the output of the optical correlator agreed with the electronics and the trained kernel is properly applied to the input image in the optical

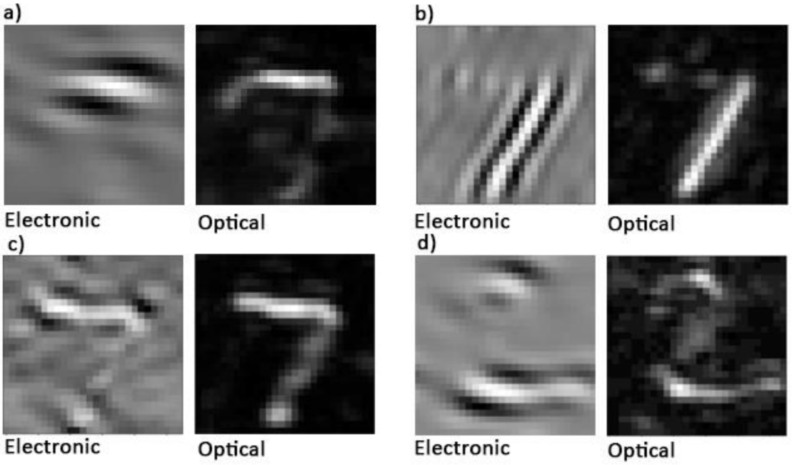


Fig. 5.6 Processed input images with electronic computing and optical 4-f optical correlator, with different kernels applied to image of digit ‘7’ and ‘2’. [109]

system. The overall tone of the convolution results of the 4-f correlator is darker, most likely because the camera captures the irradiance rather than the amplitude of the light.

Fig. 5.7 shows the confusion matrix of the CNN inference with a subset of roughly 300 images of the MNIST dataset. The overall accuracy is 81.01% which is lower than the accuracy of an electronic CNN (98.2%). Subsequent tests using the same MNIST subset resulted in similar overall accuracies (±1%). The matrix shows that most digits were correctly identified with high accuracy. A clear exception is the digit 1, which only has an accuracy of 14.0%: This may be due to the misalignment of the camera which causes normally inactive neurons in the FC layers to be activated for this few-feature digit. Also the chance of FC layer overfitting is high, since they are only trained with ‘perfect’ electronic convolution. Furthermore, the lower overall accuracy might be caused by flickering, which is caused by the asynchronous refresh rate of the camera and SLMs. If we disregard the entries of the 1-row, assuming that these are inaccurate due to misalignment or flickering, the accuracy is 91.57%. This is similar to the results with DMD devices [102]. Due to the lack of light intensity and to minimize flickering, the exposure time of the camera, and thus the convolution time of the correlator was fixed to a rather high of 0.3s, but the time could be reduced significantly by using faster, more sensitive cameras.

Furthermore, it would be interesting to see what accuracy could be achieved while exploiting maximum parallelism by displaying a grid of input images on SLM1. Additionally, the accuracy that was achieved in this work is likely lower

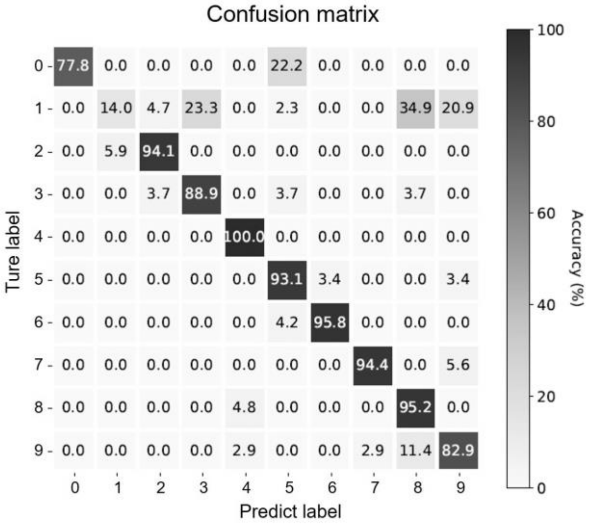


Fig. 5.7 Confusion matrix of the inference on the optical correlator, showing accuracy [109].

than the electronic accuracy because of misalignment, flickering and optical aberration. However, the 4-f optical correlator based CNN system built with low-cost components, such as amplitude modulated SLM, and simple convex optical lens and conventional camera, has shown similar performance respecting to the state-of-the-art DMD-based CNN system.

In future work, since the FC-layer after the convolutional layer is implemented in the electronics, which is not further optimized in this experiment. If the FC layers could be further trained and optimised with the outputs of the optical correlator, and the network can adjust to slight misalignments in the optical correlator, an improvement of the performance of the 4-f CNN system is expected. The flickering issue can be addressed in future work by utilizing a camera and SLMs which support generator locking (Genlock). Finally, considering the diffraction orders of the used light modulators when choosing their lenses is expected enhance the light modulation, since the diffraction orders can result in unwanted interference for large input matrices.

5.3. Photonic CNN with SOA-based Cross-connect

5.3.1. Structure and operation of PCNN

The cross-connect chip described in Chapter 2 integrates up to 8 linear neurons per layer for deep neural network. In this section, the PCNN implementation with the SOA-based cross-connect will be explained and discussed.

Fig. 5.8 represents the structure of the cross-connect chip, with the indication of the operation units for the implementation of PCNN. The three stages presented in Chapter 2 are now exploited as pre-amplifiers, filter selection, and filter stages. The most important stage is the filter stage in the cross-connect, which includes 8 filters in a column. Since there are 8 weighting SOAs per filter, the weighting elements available now on chip is 8. This means that such filter can perform or a 2×4, or a 4×2, or 2 times a 2×2 kernel matrix calculation, for a total of 64 element-wise multiplications (as we have 8 filters in a column). The 2×(2×2) matrix multiplication can be employed for a filter multiplication as shown in Fig. 5.3, where the filter includes 2 times 2×2 kernel matrices. While the weight SOAs are set on individual input wavelength for a WDM input corresponding to the input matrix, with each lambda corresponding to a pixel. The scanning of the input images as shown in Fig. 5.1 can be realized in the spatial and time-domain, with amplitude modulated input lambdas at high speed, e.g., a 10 Gbaud/s, as used in Chapter 2. The operation scheme with WDM and time multiplexing is explained in the next section.

Since the number of the kernel elements is fixed for the cross-connect chip. one can defined the operation scheme of the convolution layer augmented with 3 dimensions: the wavelength, the time and the space domain, or in other words

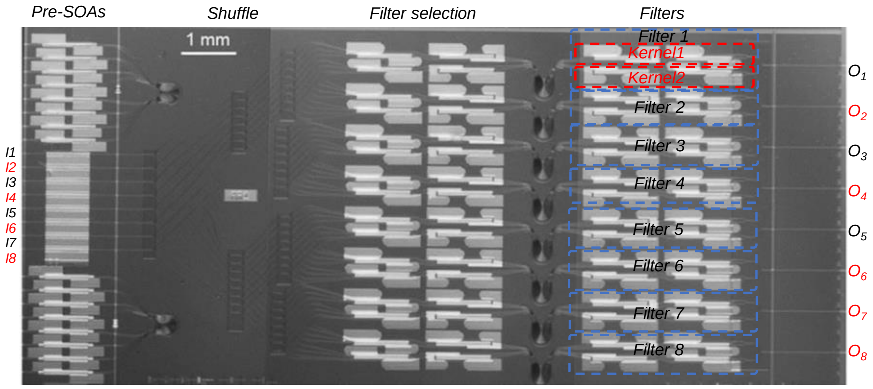


Fig. 5.8 The cross-connect chip with notation for convolutional layer operation.

with WDM and time-multiplexing, and multiple inputs utilized for spatial domain operations.

5.3.2. AWG-SOA-Coupler filter structure

AWG-SOA-Coupler with exploitation of a single FSR. Fig. 5.9 illustrates the encoding scheme of the input image with 8 × 8 pixels with 9-bit precision, as done in Fig. 3.3, chapter 3, and the operation of the first two AWG-SOA-coupler filters (black dashed line) within the cross-connect chip . The gray images at the input represent 2-dimension input images, which represents two different dimensions of the same input image, D_1 and D_2. According to the kernel size (2×2), the input images is sliced into fractions of 2×2 pixels images, where the different colour on the image represents 4 wavelengths (4 pixels) for input I_1. The information of the input images in one time slot for D_1 is encoded in λ_1, λ_2, λ_3 and λ_4, shown as red, yellow, orange and green, respectively. And the information of D_2 is encoded in λ_5, λ_6, λ_7, and λ_8, as shown at the bottom left. The pixels in rows of input images are flattened into serial sequence in time domain as in [113], shown as time slot of t_1, t_2, t_3, and t_4, where for the I_1 and I_2 is highlight with dark colour on the first rows in images D_1 and D_2. And serial sequences for I_3 to I_8 is shown with lighter colour for the other three rows at the same images. To clarify the representation, the stride of the scanning of the kernel in time domain is set to be the same as the width of the kernel, and the stride scanning in the spatial domain is set to be the same as the height of the kernel, which is 2 in this case. This operation reduces the size of the input image from 8×8 to 4×4 for the output after the filters. The detected output O_1 and O_2 in the time window t_1 will be the first element of the output matrix for filter 1 and 2, respectively.

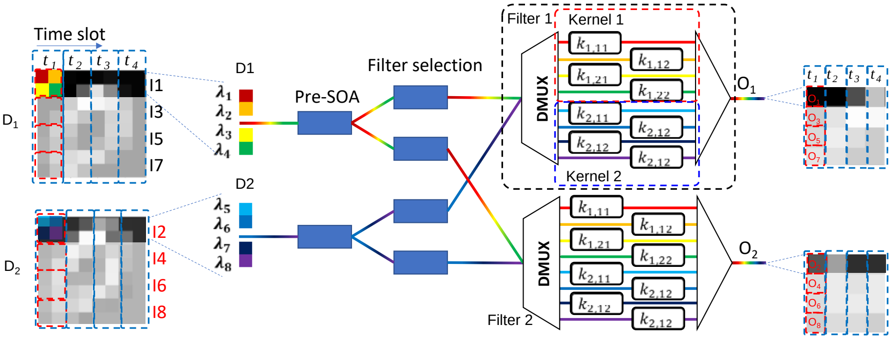


Fig. 5.9 The schematic of the operation for one unit of the optical convolutional layer, including 2 filters with 2 × 2 kernels.

Similarly, the cross-connect consists of 4 times the data input units as in Fig. 5.8, using 2 inputs every 2 data (dimension) inputs, for a total of 8 inputs. The encoding for these ports is also shown at the input 2-dimension image, in the spatial direction from the top to bottom, by denoting them as I_1 up until I_8. As the pair I_1 and I_2 provide O_1 and O_2, the I_3-I_4, I_5-I_6, and I_7-I_8 pairs will provided the outputs pairs O_3-O_4, O_5-O_6, and O_7-O_8, respectively. As indicated at the output in the figure, the O_1 to O_4 outputs represent the 4×1 vector of the first column for the filter 1, and the O_5 to O_8 outputs give the first column for the filter 2. Hence, for one time slot, one can obtain two 4×1 vectors at the output simultaneously for filter 1 and filter 2. Note that the kernels in filter 1 and filter 2 is replicated for the other parts (filter 3 to 8) of the chip as indicated in the Fig. 5.8, which suggests that the potential of the 64 weight SOAs in the filter stage is not fully exploited for parallel computing. This is due to the couplers used at the outputs of the chip, which combine all the weighted addition from the filter and give only one output for one filter.

The spatial domain and WDM is employed in the parallel computation of element-wise multiplication. In the next section, the dimension of using free space range can be applied to the exited structure on chip to accelerate the computing, by adding FSR filters at the output.

AWG-SOA-Coupler with exploitation of Multi-FSRs. To further increase the parallelism of the convolutions, using the same kernel to process different (scanned) data, one can exploit the free space range (FSR) regions of an AWG. Fig. 5.10 depicts the operation with additional FSR regions of the same AWG: by using 2 of the FSR regions (denoting this with FSR = 2), one weighting element can set a weight factor to 2 wavelengths in two different FSRs, and at the output of the chip, using FSR filters, the output can be demultiplexed into different FSRs

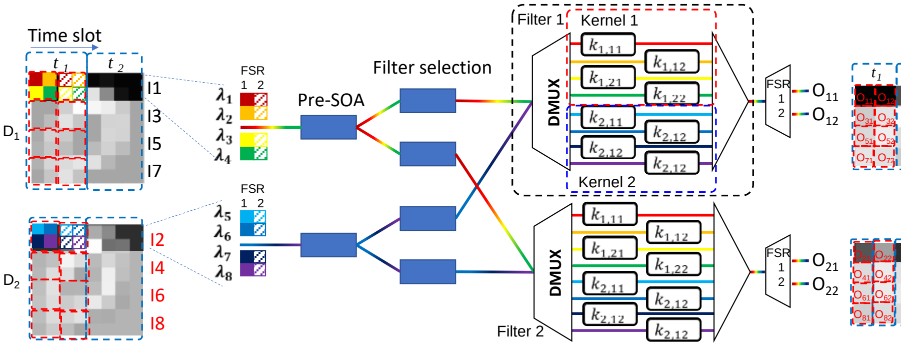


Fig. 5.10 The schematic of the operation for one unit of the optical convolutional layer, including 2 filters with 2 × 2 kernels, with 2 FSR, using 16 wavelengths.

to achieve FSR separation and providing additional output elements. By implementing FSR = 2, the input images can be encoded 2 times faster as it is about encoding 2 times more matrices, therefore the original t_2 time slot will be replaced with the wavelength in 2nd FSR, and the computing speed will be doubled, compared to the operations in Fig. 5.9.

Here, the AWG-SOA-coupler filter with multi-FSR exploits the parallelisms in of space, wavelength, FSR region, as the same as the state-of-the-art PCNN implementation with Xbar-PCM. If the filter structure is improved, one can define two more implementations with additional operation domain, which will be explained in the next section.

5.3.3. AWG-SOA-AWG filter structure

AWG-SOA-AWG with exploitation of a single-FSR. In order to further increase the wavelengths sent to kernels, an improved filter design utilizing AWG-SOA-AWG structure with mirrored cyclic AWGs at the output [108], instead of couplers, can process signal from multiple input ports and provide summation on mirrored multi-output ports after one filter. Fig. 5.11 illustrates the improved operation scheme of PCNN with the 3-stages cross-connect and replacing the output couplers with mirrored 8-channel AWG as a multiplexer, providing 8 outputs from 1 filter, with each output giving a summation of two 2×2 matrix and 2×2 kernel multiplications. In this case, the input image from different channels is sliced into 2×2 matrices in the same way as the previous operation: the two input 2×2 matrices from D_1 and D_2 are encoded into 8 wavelengths, from λ_1 to λ_8. However, for this case, the D_1 and D_2 images are input to the photonic processor at the same port, so that the input I_1 includes λ_1 to λ_8 with the information from the 2-

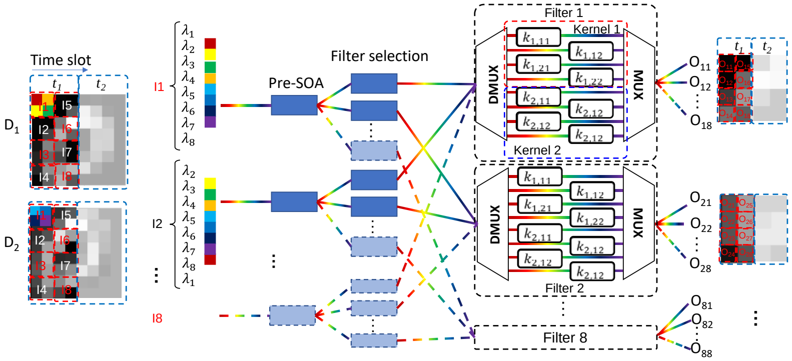


Fig. 5.11 The schematic of the operation for improved design of the optical convolutional layer, including 8 filters with 2 × 2 kernels for 2-channel input image, with single FSR and cyclic λ of AWGs.

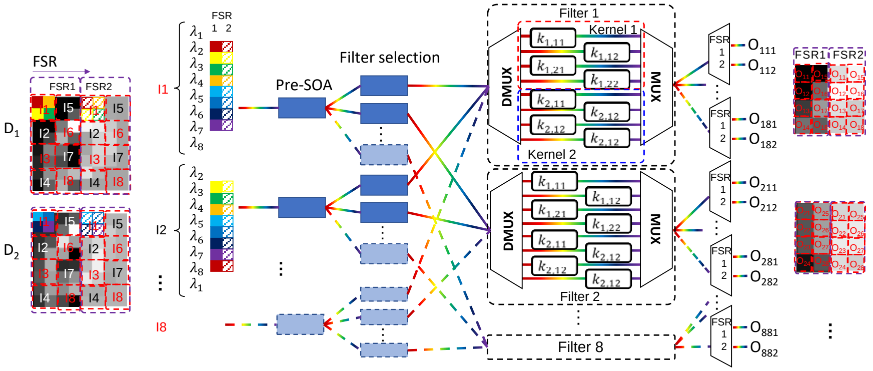


Fig. 5.12 The schematic of the operation for improved design of the optical convolutional layer, including 8 filters with 2 × 2 kernels for 2-channel input image, with FSR = 2 and cyclic λ of AWGs.

dimension input images. And because of the cyclic property of the AWG, the information of I_2 should be encoded into λ_2 and go to λ_8 and λ_1, with one wavelength shift to correctly localized the pixels to the same filters. In this way, the first element of the kernel 1 in filter 1 will process the *i*-th wavelength λ_i from the *i*-th input, i.e., the same weighting element process 8 wavelengths from different inputs, which provides a simultaneous scanning in the spatial domain. And the operation will be similar for the other filters. Then at the output mirrored AWG, the 8 wavelengths from *i*-th input will be focused back to the *i*-th output as the summation of element-wise matrix multiplications of the information of the *i*-th input, resulting in a 4×2 matrix for one filter. The cross-connect chip with 8 filters can process different kernels in this setting, instead of replicating the filters as shown in previous operation. As indicated in the figure, for one time slot t_1, shown in the blue dashed box, the 8 inputs include 2-channel 8 times 2×2 matrices, the complete convolution can be achieved for two time slots for a stride step of 2.

Up to now the parallelism improved by the addition of the wavelength domain, spatial domain and cyclic wavelength domain is considered. But now, we can go even further to includes the different FSR regions for computing.

AWG-SOA-AWG with exploitation of Multi-FSRs. Similar to the operation in Fig. 5.10, the wavelengths used for the encoding cover 2 FSR regions, resulting in 16 wavelengths: the encoding of input matrix in the 2nd time slot is replace with the 2nd FSR, as shown in Fig. 5.12, so that with the FSR demultiplexers at the output, the output elements for forming the output matrix are doubled. In this case, the same weighting elements will process 2 FSR of cyclic wavelengths, which accelerates the computing by a factor given by the FSR number considered.

In the next section, the convolutional operations with the mentioned structures will be explained mathematically, and the performance of the cross-connect chip will be discussed in terms of the computing speed, power consumption and compared to the state-of-the-art PCNN.

5.3.4. The performance of SOA-based PCNN

For the first PCNN operation shown in Fig. 5.9, the i-th input I_i is a $k\times k$ matrix X_{mj}, which is the j-th matrix of the m-th channel input image. X_{mj} for $m = 1, 2, \ldots, M$, can be encoded into the same input port with different wavelengths or separately encoded into M input ports for different input channel images for kernel size of $k\times k$. The Fig. 5.9 shows the latter case with separated ports for different channel images. In this case, for N $(N=4)$ times spatial scanning of the kernels, with j = 1, 2, …, N, so that $j = (i+1)/2$ for i is odd, and $j = i/2$ for i is even, $m = 1$ and 2 for images in D_1 and D_2. The l-th filter includes M kernels for given M-channel input images, with $l = 1, 2, \ldots, L$. With m-th kernel denoted as K_{lm} in the l-th filter, the j-th output for the filter is:

$$y_{lj} = \sum \mathrm{K}_{lm} * X_{mj} \tag{5.2}$$

with * denoting the element-wise multiplication. Therefore, for a given N spatial scanning, the total number of elements from one filter will form a N by 1 vector for 2-dimesion input images. This is done by replicating the input format and the L-filters to other aera of the chip for N times, which is limited by the input ports available on-chip. For this operation, with $k\times k$ kernels, M-channel input image, it requires $k\times k\times M$ weighting elements in one filter. Since the multiplication of the kernel with the input matrix is done in single wavelength, the required number of wavelengths is also $k\times k\times M$.

If further exploit R times FSR of the AWG as shown in Fig. 5.10, providing additional wavelengths for the input encoding, the $k\times k$ kernel can process $R\times k\times k$ wavelengths. The single output elements will be:

$$y_{ljr} = \sum \mathrm{K}_{lm} * X_{mjr} \tag{5.3}$$

for the r-th FSR of the j-th output from the filter for the l-th filter. In total, the cross-connect structure can process $R\times k\times k\times M\times L\times N$ times element-wise multiplication and provides $R\times L\times N$ output elements, in which $R\times N$ for one filter. For the case of using the recent chip as shown in Fig. 5.8, with available FSR of 2, the total element-wise multiplication is 128 and it provides 16 outputs for single time slot.

On the other hand, when replacing the output couplers with the mirrored AWG, as shown in Fig. 5.11, the encoding scheme is slightly changed as mentioned in the previous section. The input matrix X_{mj}, the j-th spatial scanning matrix of the m-th channel input image, is encoded into input i, this time, with $i=j$, and using in total $k\times k\times M$ input wavelengths. The inputs use the same number of input wavelength but encoded with one wavelength shift for the other input ports. The single output is also defined by equation (5.2). However, the $k\times k\times M$ weighting elements process $k\times k\times M\times N$ element-wise multiplications since there are N cyclic wavelengths in a single weighting SOA, the N scanning of the input images is realized by using N wavelengths in N input ports, instead of $k\times k$ wavelengths for the case of using couplers. Hence, N will be the same as the total number of weight elements enabled by one AWG, i.e., $N = k\times k\times M$ in this case. The improved operation structure can also utilize the FSR of the AWG. If R times FSR wavelengths is enabled for the kernel operation, as shown in Fig. 5.12, one weighting elements will process $R\times N$ wavelengths, and the $k\times k$ kernel can process $R\times N\times k\times k$ wavelengths, resulting in again $R\times k\times k\times M\times L\times N$ element-wise multiplication, and $R\times L\times N$ times summation. Note the L and N are different from the case when using output couplers. For the time being, N is limited by the available ports of the AWGs, with is defined by the weighting element number. Using the mirrored AWGs at the outputs and exploiting FSR = 2, together with the FSR demultiplexers after each output of the AWGs, the recent weighting 64 SOAs on-chip is capable to carry out 1024 element-wise multiplications and provide 128 outputs, respectively, with output elements defined in equation (5.3). Furthermore, the FSR used in the processing is defined by the bandwidth of the SOAs and the channel spacing of the AWGs, e.g., using the same SOAs in the cross-connect, by decreasing the channel spacing from 400 to 100 GHz, the usable FSR will change from 2 to 8, with unchanged number of weighting elements, the processing speed will be 4 times higher, albeit requires FSR filters for 8 FSR outputs. And for a larger kernel multiplication, a bigger N is needed, e.g., for a 4×4 kernel for 2-channel input images, it requires 32 input wavelengths, which can also be obtained by reducing the channel spacing from 400 to 100 GHz. However, the weighting SOAs in one filter should also increase from 8 SOAs to 32 SOAs, and the FSR will be kept being the same as 2, due to the limited bandwidth of SOAs.

In the following, the potential performance of the SOA-AWG based PCNN as well as the free-space optics-based CNN are compared with the state-of-the-art PCNN processors. For a fair comparison, the same dimension of the input images, of filter and of kernels, are considered: here 2-channel input image of 8×8 pixels, and 2-filters with 2 times 2×2 kernels are considered, and therefore, the weighting elements must be 16 in number. The TABLE 5.1 lists the performance of different PCNN in terms of parallelism, computing speed, energy efficiency, footprint, reconfiguration time. One can note that the integrated PCNNs have advantage over

TABLE 5.1. comparison of state-of-the-art PCNN with SOA-AWG approach

PCNN	Domain	Modulation speed	WDM Inputs	Speed* (GMAC/s)	Speed** (10GHz inputs) (GMAC/s)	Energy Efficiency** (pJ/MAC)	footprint (μm^2)	Reconfiguration time	Multi-filter summation (domain)	ref
4f-DMD	Spatial	1kHz	1	16.7	-		-	1 ms	Electrical	[102]
4f-SLM	Spatial	240Hz	1	4.0	-		-	4 ms	Electrical	[109]
MRR-PD	Spatial, wavelength	5GHz	4	80.0	160	3.2	50×50	20μs	Electrical	[105]
Modal-PCM	Spatial, wavelength	1kHz	4	16.0×10^{-6}	160	-	250×250	24s	Electrical	[106]
Xbar-PCM	Spatial, wavelength, FSR	2GHz	16	64	320^{***}	-	250×250	1ms	Optical	[108]
AWG-SOA-coupler	Spatial, wavelength, FSR	10GHz	16	320	320^{***}	2.1	500×150	5ns	Optical	S.5.3
AWG-SOA-AWG	Spatial, wavelength, FSR, cyclic wavelength	10GHz	16	2.56×10^{3}	$2.56\times10^{3***}$	0.26	500×150	5ns	Optical	S.5.3

* Calculated with the experimental modulation speed shown in the 3rd column, and with 16 weighting elements.
** Calculated with the same input modulation speed of 10GHz for fair comparison.
*** Considering available number of FSR is 2, and 8 wavelength each FSR region.

the free-space CNN, which is attributable to the fast modulation of the optical input. From the parallelism used in the PCNN, we can classify four different dimensions: space, wavelength, FSR and cyclic wavelength. Increasing the parallel dimensions will result in multiple-times faster computing, given the same weighting elements. The spatial multiplexing is used to generate parallel outputs for spatial scanning of the kernels as shown in Fig. 5.9 to 5.12, and it can be commonly used for all the integrated PCNN. Within one spatial input, the other dimension of input will play an important role to increase the simultaneous computing with one weighting element.

For 16 weighting elements, the WDM will provide 16 MAC operations due to the individual weighting. If the FSR is not used and the summation of the result of filters is summed up electrically, the required WDM input is 4 wavelengths, for instance, implemented with Microring-PD [105] and Modal-PCM [106]: the computing speed for this case is 160 GMAC/s for two-filter implementation. If the FSR is implemented with additional FSR demultiplexer at the end of the chip, e.g., the cross-bars with PCM (Xbar-PCM) [108] and the AWG-SOA-coupler structure as descripted above, the computing speed is 2 times the 160 GMAC/s, resulting in 320 GMAC/s, with used FSR region of 2.

Furthermore, if the FSR is considered and also a cyclic wavelength shift is utilized, the 16-wavelength input (with FSR of 2) can be cyclically fed to the same weighting elements to enable parallel computing, e.g. AWG-SOA-AWG structure with FSR filter at the outputs: therefore, the computing speed for this case is 8 times the 320 GMAC/s, resulting in 2.56 TeraMAC/s for two filter outputs. Finally, taking the spatial domain into account, with 64 SOAs on one chip, the cross-connect structure will be capable to process 10.24 TeraMAC/s, with the structure explicitly explained in Fig. 5.13. As we mentioned above, with

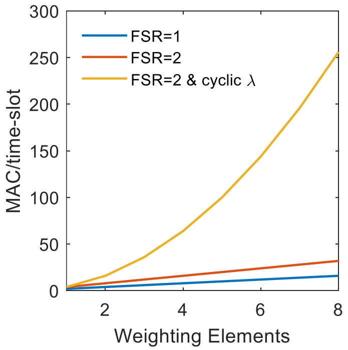


Fig. 5.13 the computing speed in MAC per time-slot when increasing the weighting elements, for scaling with single FSR (FSR = 1) (blue), with FSR = 2 (red), and with FSR = 2 and cyclic λ (yellow).

increasing the FSR region number from 2 to 8 and replace FSR filters accordingly, the 64 SOAs on the filter stage can process 4 times of 10.24 TeraMAC/s, resulting in 40.96 TeraMAC/s computing speed. Fig. 5.13 plots the number of operations per time slot for 2-filter outputs as a function of the number of weighting elements from 1 to 8, for the WDM input cases: with single FSR (FSR = 1), with double FSR (FSR = 2), and with FSR = 2 and cyclic λ. The number of FSR = 4 is considered for the latter two cases. one can see the operation number per time slot is linearly scaling for the first two cases, while it is quadratically increasing when both a higher number of FSR and cyclic λs are considered.

Finally, owning to the increase of the computing speed of the PCNN processor with multiple novel parallelism enabled by the cyclic wavelength input, the energy efficiency of the cross-connect is calculated to be 0.26 pJ/MAC for a 10 GBaud input, which is 4 times higher than the MRR-PD PCNN, while the computing speed results to be N ($N = 8$) times faster than the state-of-the-art operation scheme with Xbar-PCM. Moreover, the reconfiguration time is at least 3-order magnitude faster than the thermal-optic tuning in MRR-PD, and 6-oder magnitude faster than the PCM tuning in Xbar-PCM, though the size of the weighting element is 2 time larger than the one of the Xbar-PCMs and 30 times larger than the one of microrings. Only moving to InP-based nano-photonics we may achieve smaller form-factor engines.

5.4. Conclusion

In this chapter, the photonic implementation of convolutional neural network is investigated. The 4-f optical correlator system can enable $O(1)$ computation complexity for the convolution operation. The experimental demonstration shows that the system can classify a subset of the MNIST dataset with an accuracy of 81.01% (or 91.57% excluding the ‘1’-row). The main sources of error are likely misalignment, flickering and optical aberration. The speed of the optical correlator is limited by the speed of the SLM and camera at the setup. When adopting higher-speed SLM and camera and exploiting spatial encoding of the input images, the 4f-SLM PCNN can process at a speed of 4.0 GMAC/s, higher than a conventional GPU (GPUs in fact computer compute convolutions with an $O(n^2)$ complexity). We further discuss the operation of the integrated PCNN utilizing the 3-stage cross-connect. The possible parallelism of light is identified in space, wavelength, free space range region and cyclic wavelength domains. Exploiting multiple parallel computing domain in convolutional operation can accelerate the speed of computation. The total computing speed is proportional to the available spatial, wavelength and FSR numbers, while it is quadratic with the cyclic wavelength numbers. This additional parallelism provided by the mirror-paired AWGs can potentially enable ultra-fast computation in CNNs. With 10 Gbaud/s input modulation speed, the 64-SOA filter stage is capable to process 16 WDM inputs from two FSR in a speed of 10.24 Tera MAC/s, which is 8 times faster than state-of-the-art photonic integrated cross-bar architectures, while allowing for a reconfiguration time in the nanosecond range. The additional dimension of parallelism paves the way to ultra-fast photonic integrated CNN accelerators.

Chapter 6

All-optical Photonic Neural Networks: alternative 2D and 3D approaches

In this chapter, the alternative approaches for all-optical DNN implementation will be introduced, talking first about a novel coherent structure designed for an InP membrane on silicon (IMOS) platform in section 6.1. The proposed all-optical neural networks, on InP generic and on the IMOS platform, are then benchmarked with the state-of-the-art electronics and photonics, in section 6.2. in terms of energy efficiency, computing density, computing speed, dynamic range, etc. Finally, in section 6.3, a perspective neural network based on 3-dimesional (3D) integrated photonic neurons is proposed, discussed and identified to be a plausible a future implementation to further enable ultra-compact, energy efficient and ultra-fast computing.

6.1. All-optical Neural Network on IMOS platform

The InP membrane on silicon (IMOS) integration is a novel integration technology which enables passive-active co-integration of photonic components with enhanced optical confinements by inserting low refractive index buffer layer between the InP membrane and a Silicon substrate [114]. This platform allows passive components with size as small as the one on the SOI platform. These are bonded together with active elements, which perform similarly as the more mature InP platform [37]. Although the coherent approach shown in [39],[55] are phase noise sensitive, the MZI-mesh based matrix multiplication unit can implement positive and negative weighting using the phase of the signal, and the footprint of the MZI is smaller than the SOA. However, there is no all-optical integration neural network in the coherent passive approach to further accelerate the computation. In this section, the implementation of all-optical neural network on IMOS platform will be discussed in terms of linear and nonlinear unit, by exploiting the coherent approach. For this reason, layers of special MZIs are used and designed as explained in this section. Within this thesis, these functionalities have been proposed, designed and realized on chip to verify the process feasibility. Nevertheless, a performance projection is already made for benchmarking.

6.1.1. MZIs with novel crossing-based thermo-optical phase shifter

The size of the weighting elements is an important parameter to consider, to increase the footprint efficiency of a neural network on chip, i.e., to increase the number of operations per unit area and to enable higher MAC operation per second. The size of matrix multiplication unit with the conventional thermo-optic phase shifter on SOI chip as shown in [39] is of order of ~250×100 µm^2, mainly attributable to the long length of phase shifters of~150 µm. The conventional phase shifter used on IMOS platform is also ~200 µm in length, which is limiting the number of weighting elements for integration. Recently, Wang et al. from the Photonic Integration (PhI) group at TU/e have developed a compact thermo-optical phase shifter based on the use of an additional doped layer on the InP membrane waveguide and designed as a crossing, with a size of 15×17 µm^2, for up to 1 order of magnitude reduction in size. Fig. 6.1*a* shows the mask details of the crossing-based phase shifter and the a micrograph of the fabricated elements. With a 4 mA driving current on the phase shifter, the small unit can achieve a π phase shift. By using 2 of these phase shifters, a 2π phase shift is possible. For the implementation of MZI used in the coherent approach, one π phase shifter and one 2π phase shifter are required [39],[115] to build a 2×2 unitary matrix multiplication unit. Fig. 6.1*b* illustrates the design of the unitary matrix multiplication unit with the crossing-based phase shifter and its micrograph. The direction of the input is different from the one in [39], which feed to the tunable coupler first then connected to the phase shifter. However, the matrix implemented in this case is the same [115], [116] and in this configuration the phase matching for multi-

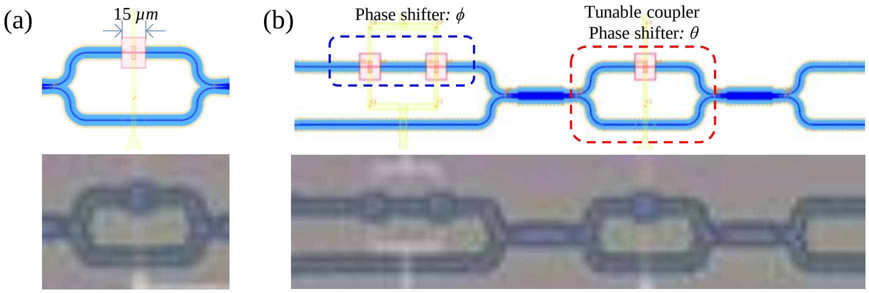


Fig. 6.1 (a) Scheme of the cross-based thermo-optic tuning phase shifter on IMOS and the fabricated elements. (b) Scheme of the Mech-Zehnder interferometer and its micrography.

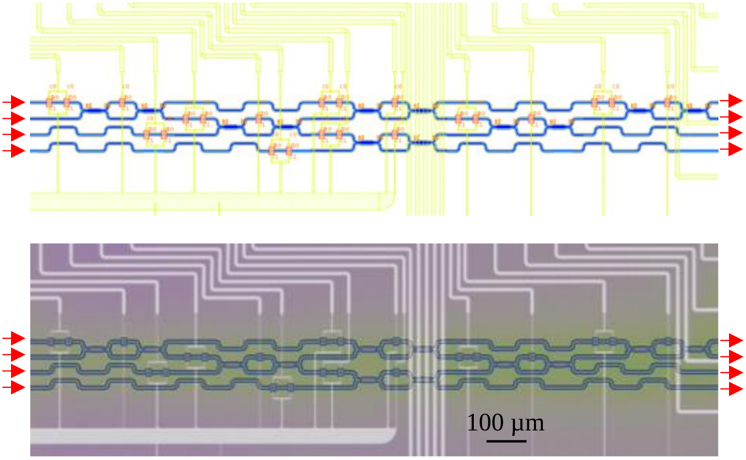


Fig. 6.2 Scheme of the 4×4 unitary matrix multiplication and its micrography.

inputs will be easier in practise, when the other direction (using phase shifter first then a tunable coupler).

The matrix multiplication is the result of multiplying the transmission matrix of the concatenated phase shifter (in blue box) and the tunable coupler (in red box) [116]:

$$T(\theta,\phi) = i\begin{bmatrix} e^{i\phi}\sin\frac{\theta}{2} & \cos\frac{\theta}{2} \\ e^{i\phi}\cos\frac{\theta}{2} & \sin\frac{\theta}{2} \end{bmatrix} \tag{6.1}$$

where $\theta \in [0,\pi]$ and $\phi \in [0,2\pi)$. The structure of the MZI mesh can be triangle or rectangular, both structures consist of $N(N-1)/2$ times of MZIs. The result of the matrix multiplication will be the multiplication of the rank-extended (with 1s filled in the diagonal elements to N rank) matrix with the effective elements shown in equation (6.1) [115]. And the final N×N unitary matrix multiplication can be implemented with a meshed network of the MZI structure. Fig. 6.2 shows the coherent MZI mesh for 4×4 unitary matrix multiplication and its micrograph. The occupation area of the recent fabricated optical inference unit on IMOS is 30% smaller than the one in [39]. The distance between the phase shifters on chip is double the values than it is possible to fabricate, and with longer waveguides overused. When using shorter connections between two phase shifters while keeping the same performance, the mesh unit can be 50% smaller, resulting in > 60% smaller than the matrix multiplication unit in [39].

6.1.2. All-optical nonlinear function on IMOS

All the optical neural networks based on MZI do not yet include the all-optical nonlinear function on-chip, which may hinder the high processing speed at the electrical connections. Recently, nonlinear functions for optical neural network implementation are proposed to be integrated on-chip, including the electric-optical and all-optical nonlinear function. The nonlinear function with electric-optical conversion may be limited by the bandwidth of the electric wires, and the impendence matching between the photodetector to the modulators needs careful design and this will require bigger footprint to accommodate the electric connections. An all-optical nonlinear function without electro-optic conversion is preferable to be integrated with the matrix multiplication unit. The optical nonlinear function based on mirroring and MZI structure is a good candidate to be integrated with the MZI mesh based optical coherent interference unit. This nonlinear function is implemented on SOI platform recently being utilized as optical thresholder [117] for optical communication. The nonlinear function is realized with the Fano resonances, combining resonances from MRR, based on power dependent free-carrier dispersion effect, and background resonates from MZI, with triggering power at 5 dBm. The triggering power of the nonlinear function can be reduced with material with higher two-photon absorption (TPA) efficiency and the processing speed can be improved with lower free-carrier lifetime in the waveguide. IMOS platform potentially outperforms SOI on the nonlinear effect [118], the InP membrane can provide >20 times higher TPA efficiency than the SOI, while the free-carrier lifetime is 10% of the one on SOI. Therefore, the triggering power of the nonlinear function with IMOS is potentially <-8 dBm, which relaxes the power requirement. Moreover, the triggering power can be further reduced with higher order micro-rings at the MZI arm. And the processing speed

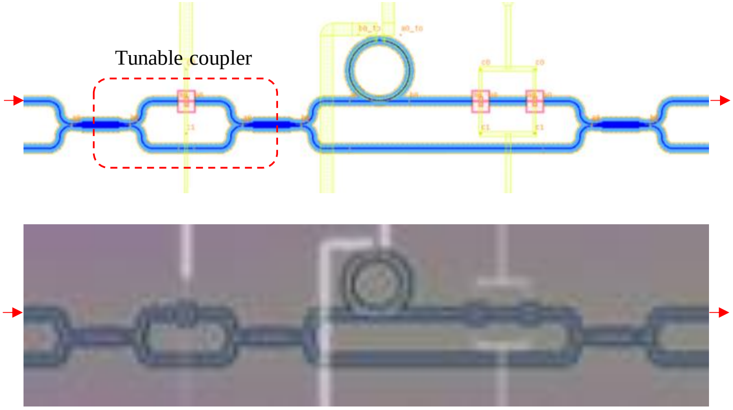


Fig. 6.3 Scheme of the MRR-MZI based nonlinear function and its micrography.

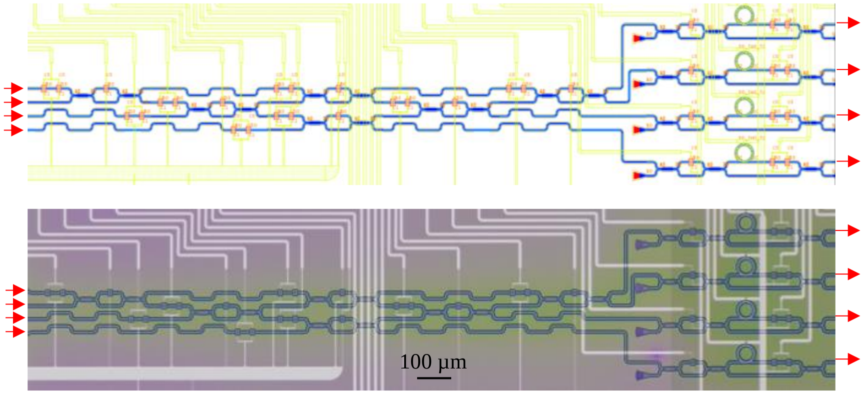


Fig. 6.4 Scheme of the 4×4 all-optical neural network layer with nonlinear function integrated and the micrography.

of the nonlinear function is potentially 10 times higher to allow processing modulated optical signal in the GHz regime, which is not easy with an SOI implementation. Fig. 6.3 shows the design and the fabricated structure of the MRR-MZI nonlinear function, which includes two parts: a tunable coupler and an MRR-MZI interferometer. The transfer function of the MRR-MZI interferometer is configured with the coupling ratio of the tunable coupler and the detuning of the MRR [117]. With a slightly detuned signal to the initial dip of the MRR, and an increase of the input power, that changes phase of arm with MRR in the MZI, both together with the changes in transmission, the sigmoid non-linear function is achievable. The size of the MRR-MZI is also reduced, compared to the SOI platform, when using the crossing based thermal optical phase shifter as in the linear part, as discussed in section 6.1.1.

6.1.3. Integrated coherent all-optical neural network

Combining the linear matrix multiplication unit and the nonlinear units, both presented above, an all-optical neural network can be established. Fig. 6.4 shows the scheme and micrograph of the fabricated triangle AONN with MZI-mesh and MRR-MZI based nonlinear function. The AONN includes 4 inputs and produce 4 outputs, which provides the nonlinear transformation of the 4×4 matrix multiplication. The fabricated AONN is expected to outperform the existed coherent neural network in terms of computing speed, footprint, energy efficiency, as discussed below.

6.2. Benchmark with the State-of-the-arts

As discussed in section 1.2, the electronic and photonic integrated circuits are developed to profoundly outperform the conventional computing structure for neuromorphic computing. In this section, the performance of proposed SOA-based all-optical neural network shown in Chapter 2 and Chapter 3 is adapted into the table of the neuromorphic implementations, for both linear synaptic operation and the nonlinear function operation.

TABLE 6.1 lists the metrics of the state-of-the-art processors designated for neuromorphic computing, compared with the SOA-based PNN processor. The electrical processers with spiking encoding, e.g., TrueNorth [14], SpiNNaker [15], Loihi [16], HiCANN [18], and Neurogrid [17], have reduced the energy efficiency down to sub pJ regime, however, the computing speed is in GMAC regime, and the footprint efficiency is at 100s MMAC/s/mm^2. The implementations with GPU-based computing structure, e.g., Google TPU [23], Nivida V100 [119], are also taking effort to optimize the structure to accelerate the matrix multiplication calculation for neural network implementation in a digital format. The energy per operation is also in sub pJ regime with the best cases, i.e., the Groq [45] or Cambricon [46], being 2-fold of the best spiking case, i.e., the TrueNorth. The computing speed of this type of accelerators are in 100s Tera MACs with optimized data flows.

In the photonic implementation, the operating frequency of the PDNN can reach 10s GHz, being 1-2 order of magnitude higher than current digital GPU, TPU, and spiking engines, leveraging WDM techniques for wavelength parallelization as an extra acceleration factor. For a 32×32 matrix-multiplication, silicon photonic DNN processor with MZIs and MRRs can provide 20 Tera MAC/s computing speed with 20 Gbaud inputs, with MAC energy efficiency in the pJ regime. The computing speed of MRR and MZI implementation is limited by the passband width of the mirroring and MZI, which is around 0.15 nm in the case presented in [35],[39], since the weighting elements is set to be static state for inference. A higher data throughput would require higher bandwidth of the filters otherwise the optical signal will be distorted [120]. This can be improved by multi-order MRRs or MZIs, however, this will increase the size of the weighting elements and may reduce the footprint efficiency. Similarly, the coherent implementation on InP-MZI in [55] and IMOS-MZI (present in section 6.1) will be also limited in 20 Gbaud input with limited passband width. The thermo-optical tuning of SOI-MRR and the SOI-MZI per weight consume 32 and 20 mW, resulting in energy efficiency of 1.6 and 1 pJ/MAC, which is the same order of the advanced electronic accelerators. The power consumption may be improved to 5 mW per weight with optimized design [34], leading to 0.25 pJ/MAC. Furthermore, using thermal tuning phase change material as in [51] will consume zero energy for inference, though the speed of the reconfiguration will be as slow as 20 µs.

In the InP flatform, the MZI with electro-optic phase shifter, with the coherent approach shown in [55], can reduce the power for the weighting to 6 mW per weight, therefore the energy per MAC is 0.3 pJ/MAC, which outperforms the all the electronics. In the recent developing IMOS platform, with a novel phase shifter based on thermal tunable heater with crossing structure, the power of each weighting MZI is 5 mW, resulting in energy efficiency of 0.25 pJ per MAC, which is 84% and 75% improvement on the SOI-MRR and SOI-MZI cases, respectively. The size of the phase shifter is as small as 15×17 µm^2, which potentially enables dense phase shifters integration. With the design shown in Fig. 6.4, the *ad-hoc* size of the MZI of 50×200 µm^2, the footprint efficiency is 4.0 Tera MAC/s/mm^2, which is half of the case for SOI-MRRs, but 2 folds of the SOI-MZI case. As mentioned in section 6.1, the sizes of the MZI can be reduced to half of the area which can provides the same performance. The footprint efficiency in this case is 8.0 Tera MAC/s/mm^2, which is the same as the SOI-MRR and 4-fold of the conventional SOI-MZI.

For the SOA-based matrix multiplication as shown in the Chapter 2, the passband width of the AWG filters is 1.46 nm, which is capable to process data up to 180 Gbaud input, however, limited by the response of the recent SOA of tens ps regime, to a throughput of 50 Gbaud [121]. This is a conservative estimation, though, as the input can be encoded in multi-levels, as shown in Chapter 3 which may not require high passband width. And response of the SOAs can be improved with quantum dot or additional carrier reservoir layer as mentioned in Chapter 3. Nonetheless, an integrated chip with 32×32 matrix multiplication can process with a rate of 51.2 Tera MAC/s and 0.84 pJ/MAC energy efficiency. Furthermore, as mentioned in Chapter 5, using the SOA-based photonic core as convolution kernel with additional parallelism can accelerate computing, by assuming FSR of 2 times, cyclic wavelength of 16 for the 32×32 matrix multiplication photonic core, for a computing speed that can be 16 folds of 51.2 Tera MAC/s and lead to 0.82 Peta MAC/s. Although the footprint of the photonic integrated circuits is bigger than the electronics, footprint efficiency of the photonic circuit is the same order of the best electronic implementation of Groq, since the computing speed is much higher. The footprint efficiency of the IMOS-MZI implementation is half of the case of SOI-MRR due to the ad-hoc design size, while the InP-SOA is 1/4 the footprint efficiency of the IMOS-MZI. However, energy efficiency of the InP-SOA implementation is 48% and 16% better than the SOI-MRR and SOI-MZI cases, respectively. Furthermore, the electro-optic reconfiguration of the weight elements is at least 3 order magnitude faster than the thermo-optical tuning.

TABLE 6.1. comparison of state-of-the-art neuromorphic processors with SOA-based approach on linear operation

	Name	typology	Computing speed (GMAC/s)	Energy per MAC (pJ)	Footprint efficiency (GMAC/s/mm^2)	Synaptic precision (bit)	ref
Electronic	TrueNorth	Spiking	17	0.27	2.0×10^{-3}	5	[14]
	SpiNNaker	Spiking	10^{-3}	6×10^{5}	4.7×10^{-5}	16	[15]
	Loihi	Spiking	3.8×10^{-4}	23.6	4.7×10^{-5}	9	[16]
	HiCANN	Spiking	5.0	198.4	8.1×10^{-4}	4	[17]
	Neruogrid	Spiking	2.3×10^{-4}	119	1.4×10^{-3}	13	[18]
	Google TPU	GPU-based	92×10^{3}	0.8	278	8	[23]
	Nvidia V100	GPU-based	1.56×10^{3}	1.2	307	32	[119]
	Cambricon	GPU-based	1.28×10^{5}	0.625	2.0×10^{-3}	16	[46]
	Graphcore	GPU-based	2.06×10^{2}	0.6	310	32	[47]
	Groq	GPU-based	8.2×10^{5}	0.36	1.13×10^{3}	32	[45]
	Myhic	Analog	1.0×10^{4}	0.5	27	8	[27]
Optical	SOI-MRR	Thermo-Optic	2.05×10^{4}	1.6	8.0×10^{3}	5	[142]
	SOI-MZI	Thermo-Optic	2.05×10^{4}	1.0	2.0×10^{3}	4	[39]
	SOI-PCM	Thermo-Optic	2.05×10^{4}	-	5.0×10^{3}	3	[122]
	IMOS-MZI	Thermo-Optic	2.05×10^{4}	0.25	8.0×10^{3}	4	S6.1
	InP-MZI	Electro-Optic	2.05×10^{4}	0.3	0.4×10^{3}	7	[55]
	InP-SOA	Electro-Optic	5.12×10^{4}	0.84	1.0×10^{3}	9	S2.1
	InP-3D	Thermo-/Electro-Optic	20.5×10^{6}	0.063	3.5×10^{4}	8	S6.3

6.3. 3D Stacked PNN

The key approaches for next generation optical engines will include the mixing of the best-in-class material and device platforms for both passive and active components, together with the promotion of in-memory processing where collocation of photonic memories with high-speed photonic MAC operations is pursued. We argue that the right combination of the best-in-class technologies will have to be conceived via a 3D stack integration approach (see Fig. 6.5), which sees the co-integration of zero-power synaptic operation that can seamlessly interact with the overlying electronics (top layer), of the lowest-loss routing functionality (middle layer), and of the most compact nonlinear functions for light generation, activation, and detection (bottom layer) for enabling neuromorphic photonic processor, which are scalable, ultra-compact, and energy efficient. Fig. 6.5 shows a schematic of the envisioned 3D ultra-compact neuron, where the vertical stacking of the three functional layers is underlined. The reason for suggesting a planar 3D integration scheme is to avoid any complex and unreliable hybrid integration scheme and offer instead flexibility in terms of material of choice, as well as ease of integration. In the following sections, we analyze the downside of current technologies and identify the desired performance for each layer to propose the most suitable technologies, whose combination promises disruptive deployment of optical computing on chip.

6.3.1. Top layer—Synaptic operation

In every synaptic connection, input data get weighted prior to reaching the nonlinear activation unit, meaning that an optical carrier or a modulated signal needs to be attenuated or delayed by a certain amount dictated by a weight value.

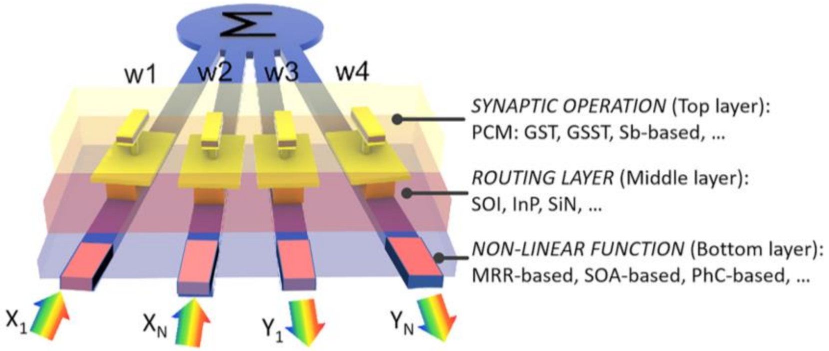

Σ
w1
w2
w3
w4
SYNAPTIC OPERATION (Top layer):
PCM: GST, GSST, Sb-based, ...
ROUTING LAYER (Middle layer):
SOI, InP, SiN, ...
NON-LINEAR FUNCTION (Bottom layer):
MRR-based, SOA-based, PhC-based, ...
X_1
X_N
Y_1
Y_N

Fig. 6.5 Schematic of the envisioned photonic neuron in a 3D fashion [3].

Photonic interferometers and resonating structures are mostly employed to realize synaptic connections, leveraging various technologies based on thermo-optic, electro-optic, non-volatile and plasmonic enhanced modulation of amplitude or phase attributes. When mapping the dynamics of the employed technology, integrated photonic weighting elements can be classified in fast and slow elements. Fast weighting solutions become essential in training routines where weights are updated on-line, while slow variants are required by layouts for inference tasks with weight assignment being carried out off-line.

Thermo-optic (TO) phase shifters loaded Mach–Zehnder interferometers or micro-ring resonators (MRRs) constitute the easiest way to realize on-chip weighting elements by changing the index of propagating modes through resistive metal wires lying on top of waveguides. In most cases, thermo-optic phase shifters are power inefficient and occupy large footprint area [123]. However, advanced fabrication techniques can lead to TO phase shifters with tens of microwatts per π phase shift efficiencies [124] while a collaborative research effort pursued within the European H2020 project PlasmoniAC attempts to de-liver sub-mW and ultra-compact thermo-optic synaptic elements via the de-ployment of plasmonics-on-SiN structures [125]. Alternatives to TO solutions refer to electro-refractive or even electro-absorptive modulators, where physical mechanisms such as the Pockels effect, Kerr effect, quantum confined Stark effect (QCSE), and free carrier modulation [126]-[130] are considered to realize fast and energy efficient weighting devices; however, a straightforward comparison between modulation principles in any layout cannot be pursued without application requirement relevance. Ultrahigh-speed modulation, for instance, can be achieved relying on mm-long $LiNbO_3$ based MZI modulators; however, their very large footprint raises concerns about scalability and yield [131]. At the other extreme, ultra-compact plasmonic modulators based on nonlinear polymers showcased modulation rates higher than >100 Gbaud, but still taking up the gauntlet of reducing excessive optical losses [132]. In fact, research groups are striving to engineer hybrid modulators employing graphene-plasmonics [133], InP membranes [134], and other emerging 2D materials [135] in an attempt to improve performance metrics in all aspects for both electro-optic and electro-absorptive modulators. Recently, indium-tin-oxide (ITO) appeared compelling exhibiting superior tunable absorption and unity refractive index change properties demonstrating low-loss operation and multi-GHz potential [136].

In the end, we may argue that the aforementioned modulation technologies are well suited for volatile based memory weighting functions, but severe challenges associated with thermal drifts and sub-optimal longevity of weight values during inference need to be overcome. Therefore, endeavors to conceive and develop photonic memristor devices gathered enormous attention, with the prospect to realize all-optical fast and zero-power nonlinear responses with long-term information retention capabilities. In particular, chalcogenide phase change materials (PCMs) exhibit strong modulation in a static, self-holding fashion, leading to ultra-compact and highly energy efficient operations. State-of-the-art demonstrations have revealed memristive behavior for chalcogenide PCMs in spiking neural networks [137] with self-holding properties and ultracompact footprint. GST ($Ge_2Sb_2Te_5$) islands deposited on top of Si_3N_4/SiO_2 waveguides formed a 15 μm long synapses element, with 3-bit precision and without any static power consumption but controlled via the repetition rate of an optical pulse. Recently, electrical switching of GST-based optical attenuators with external heaters79,80 and on-chip integrated PIN heaters81 have shown promising results, however, incurring in large insertion loss due to the use of ITO heaters or uniformly doped silicon heaters and in a number of switching cycles limited to ∼5–50. Ultra-compact hybrid Ge–Sb–Se–Te (GSST)–silicon Mach–Zehnder modulators employing an optimized electro-thermal switching mechanism [122] have also been reported. A zero-power memristive weighting structure must be developed to eliminate the energy cost of photonic linear neuron operations. However, the transmitted light amplitude should not be tuned via absorption but through phase change for removing insertion losses. Hereby, we propose some very promising novel antimony (Sb)-based compounds, which allows tuning the optical refractive index without affecting optical absorption levels. Recent fabrication ellipsometry results [82] on the non-volatile refractive index change in Sb_2S_3 (antimony trisulfide) and Sb_2Se_3 (antimony triselenide) compounds show already distinctive large index change of $\Delta n = 0.77$ without notable increase of the absorption either in the crystalline or amorphous state across a large optical spectrum of >800 nm and a switching extinction ratio with up to 5-bit resolution.

6.3.2. Middle layer—Routing layer

The right choice for a routing layer (middle layer in Fig. 6.5) is strictly connected to the architectural choice. Multi synaptic connection including the fan-in, i.e., all the inputs/outputs (I/Os) of the optical engines, construct a linear neuron. Linear neuron architectural implementations fall into two main categories: coherent and incoherent. Coherent based linear neurons, so far, relied mostly on interferometers or any type of photonic device that resembles a beam splitter arranged in mesh topologies to perform MAC operation using opto-electronics (OEs). For this case, single laser sources have been used, and fruitful interferometric layouts, supporting different cell symmetries, have been in the spotlight [37] resulting in multipath interference patterns aiming at unity fidelity features by minimizing the number of active components and mitigating phase errors. Although remarkable results have been reported so far, coherent approaches still struggle to adopt WDM functionalities and hence not fully taking advantage of the benefits of the inherently parallelized photons, which improves scalability, allowing for an all-to-all interconnectivity, and form factor. On the other hand, the first approaches for realizing neuromorphic photonic layouts relied on incoherent deployments that inherently employ WDM-enabled MAC operations, with the most recent on-chip implementations promoting the use of multiwavelength laser sources or multiple laser sources on chip and creating expectations for skyrocketing computational speed for next-gen photonic processors.

In this context, various incoherent architectures have been demonstrated spanning from crossbar arrays [108] to specialized broadcast and weight layouts on SOI [142] as well as adopting cross-connect schemes on InP [58] using WDM sources. In practice though, both coherent and incoherent based layouts face multitude performance trade-offs associated with thermal stability, increased channel crosstalk, and excessive insertion losses.

Putting facts in perspective, those trade-offs originate mostly from the waveguide platform of choice and their constituent materials. Current implementations employ heavily SOI wafers with crystalline Si waveguides to implement weighting functions and hybrid assemble fan-in structures realized exploiting multi-project wafer services or in-house fabrication facilities. High-index contrast of silicon waveguides allow indeed for compact layouts due to a high refractive index contrast at the expense of increased optical scattering and amenability to phase errors along the direction of propagation. Propagation losses in active–passive InP wafers, on the other side, are mainly due to the p-doped cladding layer: They would necessitate an *ad-hoc* process development to guarantee competitive waveguide losses [143]. In stark contrast, moderate index contrast platforms such as those based on silicon nitride propelled the deployment of photonic devices with higher immunity to the temperature drifts, lower optical losses, improved crosstalk values, and wider wavelength transparency [144]. In addition, silicon-rich nitride platforms emerged as the means to tailor the index contrast of photonic devices complying with applications requirements where increased index contrast is imperatively needed, such as low-loss (<1 dB) fiber to chip coupling using grating couplers [145]. Multi-layer silicon nitride cross-connects standing out as 10 × 100 any-to-any as well as feedforward networks have been also demonstrated pointing out the increased degree of freedom in designing scalable linear neurons pursuing low-loss SiN platforms [146]. Nevertheless, hybrid integration of active devices on SiN platforms is still in its infancy impeding their wide adoption in practical applications as opposed to the more mature approaches of co-integrated CMOS electronics on SOI. For these reasons, while we foresee that the SiN platform will play a key role as routing layer in neuromorphic photonics, we also suggest that only a 3D integration scheme merging SiN with best-in-class active technologies bear promises for a revolutionary neuromorphic photonic platform.

6.3.3. Bottom layer—Nonlinear function

After identifying the synaptic weight and linear neuron (top and middle layer), we reach the bottom layer of the envisioned 3D integrated approach in Fig. 6.5, where all the active functions (of generation, detection, and activation) must be placed and interfaced to the linear neuron to realize the complete neural units.

Nonlinear activation function (NLAF) elements can be classified in optoelectronic (OE) or all-optical (AO) ones. Regarding O/E NLAFs, the nonlinear optical transfer function is mostly mediated by an electrical signal used to convert optical input signals at the output [40],[147],[149]; however, the employed O/E conversion can increase inference latency [150] or require efficient and low power optoelectronic devices leveraging advanced fabrication methods [151]. Demarcating from O/E based NLAFs, all-optical variants are highly anticipated to revolutionize neuromorphic photonic circuits by providing time-of-flight latencies, exploiting fully the available optical bandwidth and consuming low power. All-optical SOA-MZI wavelength converters [56] have been shown to exhibit a sigmoid-like transfer function with an extinction ratio of 11 dB at 10 GHz, yet in a bulk and power hungry deployment. On the other hand, power efficient Fano-based MRR-MZI structures [51],[152]-[155] can operate as optical thresholders achieving energy efficiencies down to ~13.3 pJ/Op; however such schemes face hurdlers to facilitate WDM inputs and increase aggregated bandwidth. For example, an optical nonlinear function has been reported in Ref. [51] with PCM GST being loaded on a MRR, using a probe laser coupled into the MRR to assist generating the output pulses; however, the pulse width limits the maximum operation speed. Plasmon-assisted localization in combination with highly nonlinear material such as CdSe quantum dots [152] allow for transfer function resembling a reversed sigmoid function, with a stunning compute efficiency of 1 Tera Op/s, pitfalls though regarding high loss and the low contrast aspects. Without moving toward more exotic implementations, a more robust and suitable technology for the active layer may rely on the resonantly enhanced nonlinear response of passive semiconductors, which can be configured to generate a tailored nonlinear function, as for instance in Ref. [153] where photonic crystal Fano structures have been used determining the optimal ratio between energy per bit and speed. Superior performance can be achieved by using photonic crystals nonlinear resonators made of III-V materials. Parameters can be optimized to favor speed and/or energy efficiency [156] for the integration on a silicon photonic circuit [157].

6.3.4. 3D hybrid integration approach in perspective

So far mainly monolithic integration, where light is moving on a single guiding plane (2D), or butt coupling of diverse monolithic integrated chips have been exploited to realize combinations of diverse functionalities (e.g., linear and nonlinear functions). While the butt coupling of diverse photonic material platforms is exploited to overcome the lack of on-chip gain in fully passive platform [36], but at the expense of a complex and unreliable coupling scheme, a monolithic integration approach offers process robustness, yet preventing further scalability because of the high passive excess losses in active–passive platforms [37].

In this section, a more interesting approach is represented by the 3D hybrid planar integration, made of multiple guiding planes stacked in a 3D fashion, each with a different functionality, where a bottom guiding layer is patterned and planarized, to deposit or bond a next layer on top, which is also patterned and planarized, and so on and so forth. Such an integration scheme may provide the most compact optical neurons. Specifically, InP over SOI hybrid technology has been previously demonstrated as an extremely promising solution for future photonic circuits as it combines CMOS compatibility with the optoelectronic properties of III-V materials [158]. The two-layer structures are separated by a low-refractive index bonding layer constituted of benzocyclobutene (BCB) and SiO_2. A sub-100 nm precise alignment of the III-V nanocavities on top of SOI waveguides below is possible via multi-layer overlay and high precision reference markers fabricated at the silicon waveguide level, which guarantees a high optical coupling between the different layers in the vertical direction. Moreover, the vertical evanescent coupling between the underlying Si waveguides and nanocavities on top is a very compelling method, whose strength can be tuned at will by controlling the transverse overlap of the electro-magnetic fields distributed in each level by changing, for example, the SOI waveguide width and/or the SiO_2 intermediate spacer layer [158].

Eventually, electrically powered InP photonic crystal (PhC) nanostructures on Silicon on Insulator (SOI) waveguide platforms have confirmed their superior compactness and energy benefits over the entire spectrum of optical active elements in a rich functional suite, including as nanolasers [99], nanomodulators

[159], nanophotodetectors [160], and bistable nanomemories [161]. The manufacturing of these nanocavities on top of SOI passives is achieved through a top-down approach that bypasses sophisticated III-V regrowth steps and annealing at temperatures beyond which the CMOS back-end of line processing cannot be envisaged.

As a result, this work paves the way to a 3D hybrid technology that opens to the convergence of microelectronics and photonics for a new generation of complex optoelectronic circuits, which can serve the novel dogma of neuromorphic photonics. Transferring the InP PhC technology onto the SiN waveguide platform and engineering their rich nonlinear characteristics into a complete set of ultra-low power and ultra-small footprint computational elements can bring to verifying the potential of photonic crystal technology in the context of neuromorphic computing, once interfaced to memory-loaded linear neurons.

6.3.5. Projected performance of the 3D stacked PNN

The integration of possibly electrically controlled (via localized thermal heaters or PIN electrical junctions) non-volatile Sb-based layers onto low-loss SiN waveguides is expected to produce record-low-loss <0.15 dB non-volatile photonic structures, allowing full π-phase shifts via a very short waveguide section (<15 μm). Sb materials are expected to yield identical weight resolution performance metrics when operated in standalone waveguide configurations, but to allow up to >8-bit weight resolution levels when bringing them into an interferometric engineered layout for the highest linear neuron accuracy. On the other side, InP nanophotonic crystals (PhCs) are envisioned to be used for ultra-low-energy ultra-fast (>40 GHz) photonic fan-in, gating, and activation function units. This InP PhC technology may be transferred under (or onto) the SiN Sb-loaded waveguide platform for the most compact and complete functional set required for neuromorphic computations that, interconnected to form programmable meshes [36], are foreseen to release sub-fJ/MAC energy efficiencies.

The adoption of this technology will release a neuromorphic photonic technology that can be benchmarked along all relevant metrics with respect to the electronic AI chip state-of-the-art (see Fig. 6.6, with corresponding metrics on TABLE 6.1). Assuming a basic 32×32 neuron architecture for this 3D photonic neural network (3D PNN), we highlight how this can scale to multi-neuron photonic processors enforcing breakthrough performance at the following metrics. The operating frequency of the 3D PNN can reach 50 GHz, being about 50 times higher than current digital GPU, TPU, and spiking engines, leveraging WDM techniques

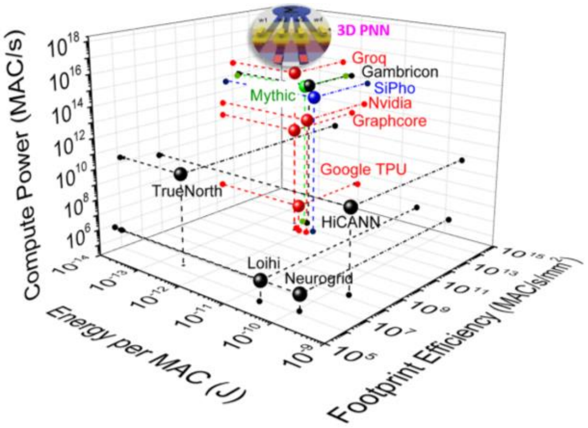


Fig. 6.6 Projected performance for our 3D photonic neural network (PNN) and comparison with state-of-the-art machines [3].

and wavelength parallelization as an extra acceleration factor. The computational power in MAC/s (or number of MACs times frequency), the case of a typical $N \times N$ crossbar architecture, with up to 32 different wavelength channels will yield tens of TMAC/s, with the total area of four-channel $N \times N$ configuration calculated to have a slightly higher footprint of ∼1 mm^2. The full-scale 3D PNN processor could encompass ∼400 cores within a standard silicon die area, yielding a total computational power of the order of tens of PMAC/s, i.e., >4× higher over the rack-scale HICANN and ∼140× more over Google's Cloud TPU v3 boards. Considering the power consumption of N InP PhC PDs, N InP lasers, and N InP modulators including the driver and a total crossbar insertion loss, the total power required for single channel operating at 50 GHz is calculated to be tens of mW. This suggests energy efficiency that scales linearly with N, leading to a value as low as ∼63 fJ/bit, i.e., improved by an order of magnitude over all available digital and analog neuromorphic engines and a total footprint efficiency that surpasses by 2 orders of magnitude all available digital, analog, and spiking engines [125]. Latency will obviously benefit from the use of photons as a data carrying medium toward time-of-flight latency values less than a few hundreds of ps even for the longest route within a 20 × 20 mm^2 silicon die. Fig. 6.6 shows the 3D PNN's breakthroughs, simultaneously achieving top-notch compute power and footprint efficiency compared to state-of-the-art machines, highlighting its long-term vision and future-proof technology.

6.4. Conclusion

The all-optical neural network design is extended to the coherent design on InP membrane on silicon platform as IMOS-MZI implementation. Utilizing a novel thermo-optic tuning phase shifter-based crossing heater, the size and the power consumption of MZI linear units is smaller than the conventional phase shifters used in the SOI platform. The nonlinear function based on MRR-MZI implemented on InP membrane waveguide potentially outperforms SOI on the nonlinear effect, with >20 times higher TPA efficiency and only 10% of the free-carrier lifetime, resulting in nonlinear triggering power < -8 dBm and enabling GHz processing speed.

The all-optical implementation proposed in this dissertation is benchmarked to the state-of-the-art electronic and photonic neuromorphic devices. The IMOS-MZI implementation shows 84% and 75% energy efficiency improvement than the SOI-MRR and SOI-MZI cases, respectively, while the footprint efficiency is the same of the case for SOI-MRRs, and 2 folds of the SOI-MZI case. For the SOA-based matrix multiplication InP platform, it shows energy efficiency of 48% and 16% better than the SOI-MRR and SOI-MZI cases due to the high passband width, while the footprint efficiency is 12.5% and 25% of case in SOI-MRR and SOI-MZI. However, with the convolutional implementation, the additional parallelism provided by the AWG-SOA-AWG structure can increase the footprint efficiency of a factor 16, which makes the footprint efficiency of the InP-SOA outperforms the other implementation. Moreover, the SOAs provide on-chip gain and higher tuning dynamic range, therefore, higher bit-precision, and the electro-optic reconfiguration of the weighting elements is at least 3 order magnitude faster than the thermo-optical tuning.

Finally, the 3D stack integrated of photonic neural network is foreseen to achieve top-notch performance in characteristic computational metrics of computational speed, energy efficiency, and footprint. The investment in novel PCM photonic memristors on ultra-loss routing platforms and the 3D co-integration with ultracompact and energy efficient InP nanophotonics, working together for the promotion of in-memory processing, are key approaches for next generation optical engines. The right combination of the best-in-class technologies conceived in a 3D stack approach will succeed in enabling fully programmable neuromorphic photonic processors with zero-power synaptic operation, ultra-low optical loss routing functionality, and a complete suite for ultra-compact nonlinear function circuitries for light generation, activation and detection, enabling disruptive deployment of optical computing.

Eventually, the development of these technologies for the release of robust technologies that allow for ultra-compact and efficient computing hardware must be pursued together with the development of the mathematical framework for neurophotonic deep learning architectures and training models, in order to actually demonstrate clear advantages in replacing or complementing state-of-the-art conventional computational approaches.

Chapter 7

Conclusions and Outlook

7.1. Conclusions

In this research work, the implementation of a monolithically integrated all-optical photonic deep neural network is explored on an InP platform, starting from the implementation of a SOA-based linear matrix multiplication unit to the co-integration of cross-gain-modulation-based all-optical nonlinear function and the scalability investigation. The photonic implementation of a convolutional neural network is also discussed to enable additional parallelism, followed by the proposal of new schemes, 2D coherent and 3D integrated, to reach high performance and smaller form-factors. Finally, the all-optical neural network implementations within this thesis are benchmarked with respect to the state-of-the-art neuromorphic networks on chip.

Specifically, in following I summarize this thesis main achievements:

Chapter 2. The SOA-based linear synaptic operation demonstrated in Chapter 2 verifies the matrix multiplication unit while the activation function is implemented via software. The on-chip SOAs are operated in the linear regime, which reduces the complexity of the weight calibration. This is demonstrated with an NRMSE smaller than 0.08 and a best-case dynamic range of 27 dB. The Iris flower classification problem is used as a small-scale example to demonstrate that the photonic neural network concept allows a similar accuracy as when using electronics. The final prediction accuracy of a three-layer photonic DNN, based on multiple conversions from the optical to the electronic domain and vice versa, implemented for solving the image classification problem, results reduced of 9.2% with respect to the simulated prediction accuracy. A comprehensive error analysis shows that the conversions (E/O and D/A and vice versa) are the main factor of signal distortion in the signal processing, which suggests that an on-chip all-optical neural network implementation is expected to outperform. The co-integration of gain elements and optical filters is foreseen to provide a large dynamic range as well as to enable the route to scalable DNNs.

Chapter 3. An all-optical neural network structure with WDM connectivity and SOA-based all-optical neurons is then proposed in Chapter 3. The linear neural network can be easily scaled as a function of the WDM signals for multi-synapsis neurons: the linear processing unit can scale up to 64 channels, while guaranteeing a large input dynamic range under neglectable error introduction. A fully monolithically integrated all-optical neuron is experimentally demonstrated by exploiting an SOA WC-based optical nonlinear function based on cross-gain modulation (XGM). The performance of the fully integrated all-optical neuron is 10% better than in the hybrid case in terms of error introduction. The all-optical neural network is simulated with noise induction for benchmarking the inference of a noisy DNN built for the MNIST handwritten digits classification problem, showing that, when working with 10 Giga sample/s inputs, the all-optical approach is about 2.5 times faster than state-of-the-art GPUs, while guaranteeing similar accuracies. Furthermore, the complete end-to-end system is evaluated by considering, in the overall system performance calculation, also the contribution of a control unit, transmitter and receiver units, together with D/A and A/D converters. The calculation results show that the effective energy per MAC operation for an all-optical connected DNN always outperforms the single-layer DNN system. Eventually, the energy efficiency results in constrains from the speed and power consumption on the electronic side, including the DAC/ADC at the transceivers and the control FPGA for the pattern generation and signal processing when we increase the number of synapses/neuron. Nevertheless, the AONN, taken individually, still performs more than two times better than state-of-the-art GPUs at the server level, excluding the energy for the cooling.

Chapter 4. The scaling of the all-optical neural network has been experimentally emulated, with photonic integrated SOA-based AON, utilizing XGM as nonlinear transfer function. The noise model has been developed for simulating the noise accumulation of the AON and for a WDM operation of the neural circuit. The model shows good agreement with the experimental data, when relying on the characteristic parameters of the used SOA components. The data are analysed to interpret the scalability of the AON in terms of input channel number (network height) and layer number (network depth). The results show that the WDM input, entering the non-linear function and realising N:1 conversion on a single output, undertake a noise compression after a certain number of layers, defining a stabilized final error at the output of the AONN. The implemented AON structure is capable of establishing a 12-input/neuron 12-neuron/layer arbitrary layer number all-optical neural network, with a final NRMSE < 0.1, with optimized input signal power at -20 dBm per channel, for a channel spacing of 100 GHz and a gain bandwidth of 32 nm. The noise model can be further used to investigate other parameters for the N:1 XGM-based conversion for optical signal processing, like noise inversion parameter, passive losses and SOA gain response, etc. Hence, utilizing WDM-input-to-single-output conversion via XGM in an SOA, the proposed AON structure can scale up to an arbitrary layer number with a large input channel number, resulting in an acceptable maximum output signal error.

Chapter 5. The photonic implementation of a convolutional neural network has been investigated in Chapter 5. A free-space 4-f optical correlator system is demonstrated to realize $O(1)$ computation complexity for the convolutional operation. The experimental demonstration shows that the 4-f system is able to solve the MNIST sub-dataset classification with an accuracy of up to 91.57%. The system can be further improved on the alignment and optical aberration. Removing the flickering of the captured images with a high-speed trigger locked camera will help to improve the performance. The speed of the optical correlator is limited by the speed of the SLM and camera at the setup: with a high-speed SLM and by exploiting spatial encoding of the input images, the 4f-SLM PCNN can process 4.0 GMAC/s, a speed higher than in conventional GPUs. We further discuss the operation of the integrated PCNN utilizing the 3-stage cross-connect chip. The parallelism of light is exploited in spatial, wavelength, free space range and cyclic wavelength domains. The total computing speed of the CNN processor is linearly proportional to the available spatial, wavelength and FSR numbers, while it is quadratic as a function of the cyclic wavelength numbers. This additional parallelism provided by the mirror-paired AWGs enables ultra-fast computation on the convolutions. With 10 Gbaud/s input modulation speed, the 64 SOA filter stage is capable to process 16 WDM inputs from two FSR for a speed of 10.24 TMAC/s, which is 8 times faster than state-of-the-art cross-bar architectures [108], while it is possible to be reconfigured even in the nanosecond range. The AWG-SOA-AWG structure paves the way to ultra-fast photonic integrated CNN accelerators.

Chapter 6. This chapter is dedicated to novel components and integration schemes proposed to improve computing performance overall as well as form factor in the specific. A 2D coherent all-optical neural network design on InP membrane on silicon platform with IMOS-MZI implementation is introduced. The novel thermo-optic tunable phase shifter-based crossing heater developed on IMOS enables size and power consumption reduction for the coherent linear unit. And the nonlinearity provided by the InP membrane waveguide makes it suitable for the MRR-MZI based nonlinear functions [117] to be implemented. This device potentially outperforms the SOI on the nonlinear effect, with >20 times higher TPA efficiency and only 10% of the free-carrier lifetime, resulting in nonlinear triggering power <-8dBm and enabled GHz processing speed. The benchmark of the IMOS-MZI and SOA-based AONN with the state-of-the-art electronic and photonic neuromorphic devices shows that the IMOS-MZI implementation is 84% and 75% more energy efficient than the SOI-MZI [39] and SOI-MRR [142] cases, respectively, while the footprint efficiency is the same as in the case of SOI-MRRs, and 2 folds the SOI-MZI case. For the SOA-based matrix multiplication on InP platform, the energy efficiency is 48% and 16% better than the SOI-MRR and SOI-MZI cases due to the high passband width, while the footprint efficiency is only 12.5% and 25% of the cases in SOI-MRR and SOI-MZI. However, with the convolution implantation, the additional parallelism provided by the AWG-SOA-AWG structure can increase the footprint efficiency with a factor of 16, which makes the footprint efficiency of the InP-SOA outperforming the other implementations.

In this same chapter, the 3D stack integrated neurons for photonic neural networks are foreseen to achieve top-notch performance in characteristic metrics of computational speed, energy efficiency and footprint. The investment in novel PCM photonic memristors and 3D co-integration with ultracompact and energy efficient InP nanophotonics, working together for the in-memory processing, are key approaches for next generation optical engines. The right combination of the best-in-class technologies conceived in a 3D stack approach will most probably achieve fully programmable neuromorphic photonic processors with zero-power synaptic operation, ultra-low optical loss routing functionality and a complete suite for ultra-compact nonlinear function circuitries for light generation, activation and detection, enabling disruptive deployment of optical computing.

Eventually, the development of these novel technologies, which allow for ultra-compact and efficient computing hardware, must be pursued together with the development of the mathematical framework for photonic deep learning architectures and photonics-aware training models, in order to actually demonstrate clear advantages in replacing or complementing state-of-the-art conventional computational approaches.

7.2. Outlook

In this dissertation, the novel SOA-based integration synaptic operation unit has been demonstrated for a classification problem. A monolithically integrated all-optical neuron with SOA-based wavelength converter as nonlinear function has been proposed and demonstrated to enable high-speed data processing and higher accuracy. The scalability of the neural network with the all-optical neuron has been investigated with the experimental emulation and noise modeling, and a noise compression effect has been shown to enable arbitrary layer depth for deep neural network. The implementation of photonic convolutional neural network has been demonstrated first in a free-space Fourier transform based system and then mapped into the SOA-based cross-connect chip, anticipating a potential acceleration of the convolution up to tens Tera MAC/s. The attempt to use MZI and MRR on the IMOS platform has been discussed to take the advantages of the novel compact phase shifter design and the potential enhancement of nonlinearity in the InP membrane waveguide.

Based on the findings from the SOA-based photonic neural network engine shown in this dissertation, the research outlook can be categorized in two main research questions: (1) How can we further increase the computing power of photonic processor? (2) How can we enable system-level implementation for further developing AI system?

For the first path, the potential investigation directions for future work to improve the performance of the scalable PNN based on the InP material platform are suggested to be the following:

Extend to the InP membrane on silicon (IMOS) platform. As discussed in Chapter 6, the IMOS platform provides higher optical confinement using the InP membrane, which will reduce the size of the optical element on-chip and keep the advantages of InP material of the passive-active co-integration. With the IMOS platform developed at TU/e, photonic neural network can be built on a monolithically integrated chip with optimized building blocks like using the generic InP platform from SMART photonics, but with a much smaller form factor and lower insertion loss components.

3-D stacked integration. As discussed in Chapter 6, the 3D integration technique, applied to state-of-the-art non-volatile weighting and routing ultra-low loss platforms as well as to ultra-compact membrane-based nonlinear nanophotonics platforms, is expected to enable high-density integration, low power consumption and high bandwidth interconnection, improving of almost two orders of magnitude the energy efficiency on state-of-the-art electronics, and achieving 3-orders magnitude higher footprint efficiency and number of MAC operation per unit time.

Quantum Photonic Neural Network. Quantum photonic technology is attracting more and more attention since quantum information processing can be carried out at room temperature, though detection is still in cryo-temperature. The superposition of the quantum states would provide high volume information process using quantum computing. The photonic neural network may be implemented as the superposition of the quantum states, which may be beneficial also for ultra-fast matrix multiplication processing.

For the second research question for future works, the recent SOA-based PNN design can be further developed toward a computing system to perform real applications. The potential investigation directions using the scalable SOA-based photonic chip are proposed to be the followings:

On-chip All-optical Neural Network. This approach relies on connecting multiple all-optical neurons and cascading multiple layers, as shown in Chapter 3, to build an all-optical neural network on a single chip. A chip with the size of the fabricated cross-connect chip of 5×5 mm^2, as shown in Chapter 2, is capable of establishing an 8 neuron/layer two-layer neural network and can be used to solve image classification problems as shown in Chapter 3. I believe this is still the way to go, opposite to the approach of realizing one single layer and rely on multiple E/O conversions.

FPGA-assisted Photonic Deep Neural Network system. As discussed in Chapter 1, Chapter 2 and Chapter 6, the SOA-based synaptic elements are capable of *ns* switching time, which can be exploited for fast reconfiguration. Although the inference is on the static trained PDNN, the fast response of the SOAs can provide time multiplexing of the processing of the PDNN when the physical size of the AONN chip is limited. With the FPGA-assisted current controller, as shown in Appendix 4.E, a PDNN system can be implemented to process optical data generated by 10 Gbit/s on-board transmitters. The FPGA-assisted PDNN system can be further investigated for the online training, which will be an interesting direction since most of the works presented in the field aims at static inference.

Photonic Convolutional Neural Network system. As discussed in Chapter 5, the additional dimension of light parallelism is identified to be the cyclic wavelength, which could enable ultra-fast convolution operation in the optical domain. Utilizing this concept to design and fabricate a photonic chip would be promising to reduce the computing complexity from $O(n^2)$ down to $O(1)$, and achieve ultra-fast convolution computing, as well. With the FPGA-based current controller, the PCNN system may outperform the recent CNN processors in computing speed and energy consumption.

Recurrent Neural Network for time-trace processing. The AONN investigated so far aims to achieve a deep learning model developed in computer science. The recurrent connection of the WDM input/output is still an open research direction, which will enable the memory effect of the neural network to process the optical time trace with dynamics in the time domain. With the noise-stable property of the multi-wavelength to single wavelength output design, the analog signal processing in the RNN recurrent neural network is also expected to be more noise tolerant and promising in terms of feedback loop number and scalability.

Bibliography

[1] Cisco white paper. https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visualnetworking-index-vni/white-paper-c11-741490.html.

[2] J. D. Kendall and S. Kumar, "The building blocks of a brain-inspired computer," Appl. Phys. Rev. 7, 011305 (2020).

[3] R. Stabile, G. Dabos, C. Vagionas, B. Shi, N. Calabretta, and N. Pleros, "Neuromorphic photonics: 2D or not 2D?," J. Appl. Phys., vol. 129, no. 20, p. 200901, May 2021.

[4] A. O. Riordan, G. Fagas, B. O'Flynn, J. Rohan, P. Galvin, and C. Ó. Mathúna, More Than Moore, International Roadmap for Devices and Systems (IRDS) white paper.

[5] K. Kitayama, M. Notomi, M. Naruse, K. Inoue, S. Kawakami, and A. Uchida, "Novel frontier of photonics for data processing—Photonic accelerator," APL Photonics 4, 090901 (2019).

[6] J. Wu, Y.-L. Shen, K. Reinhardt, H. Szu, and B. Dong, "A nanotechnology enhancement to Moore's law," Appl. Comput. Intell. Soft Comput. 2013, 426962 (2013).

[7] S. J. Ben Yoo and D. A. B. Miller, "Nanophotonic computing: Scalable and energy-efficient computing with attojoule nanophotonics," in 2017 IEEE Photonics Society Summer Topical Meeting Series (SUM), San Juan (IEEE, 2017), pp. 1–2.

[8] B. Marr, B. Degnan, P. Hasler, and D. Anderson, "Scaling energy per operation via an asynchronous pipeline," IEEE Trans. Very Large Scale Integr. Syst. 21(1), 147–151 (2013).

[9] Y. Shen et al., "Silicon photonics for extreme scale systems," J. Lightwave Technol. 37(2), 245–259 (2019).

[10] J. D. Meindl, "Beyond Moore's law: The interconnect era," IEEE Comput. Sci. Eng. 5(1), 20–24 (2003).

[11] "Top 500 List - June 2020," TOP500, June 2020, available at https://www.top500.org/lists/top500/list/2020/06/.

[12] J. Hasler and B. Marr, Front. Neurosci. 7, 118 (2013).

[13] A. Szalay and J. Gray, "Science in an exponential world," Nature 440(7083), 413–414 (2006).

[14] F. Akopyan et al., "TrueNorth: Design and tool flow of a 65mW 1 million neuron programmable neurosynaptic chip," IEEE Trans. Comput. Aided Design Integr. Circuits Syst. 34(10), 1537 (2015).

[15] S. B. Furber, F. Galluppi, S. Temple, and L. A. Plana, "The SpiNNaker project," Proc. IEEE, vol. 102, no. 5, pp. 652–665, 2014.

[16] M. Davies et al., "Loihi: A Neuromorphic Manycore Processor with On-Chip Learning," IEEE Micro, vol. 38, no. 1, 2018.

[17] B. V. Benjamin et al., "Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations," Proc. IEEE 102(5), 699–716 (2014).

[18] J. Schemmel, D. Brüderle, A. Grübl, M. Hock, K. Meier, and S. Millner, "A wafer-scale neuromorphic hardware system for large-scale neural modeling," in in Proc. Int. Symp. Circuits Syst., pp. 1947–1950, 2010.

[19] A. Tavanaei et al., "Deep learning in spiking neural networks," Neural Networks 111, 47–63 (2019).

[20] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," Bull. Math. Biophys. 5(4), 115–133 (1943).

[21] E. Nurvitadhi, D. Sheffield, J. Sim, A. Mishra, G. Venkatesh, and D. Marr, "Accelerating binarized neural networks: Comparison of FPGA, CPU, GPU, and ASIC," in 2016 International Conference on Field-Programmable Technology (FPT) (IEEE, 2016), pp. 77–84.

[22] V. Gupta, A. Gavrilovska, K. Schwan, H. Kharche, N. Tolia, V. Talwar, and P. Ranganathan, "GViM: GPU-accelerated virtual machines," in Proceedings of the 3rd ACM Workshop on System-Level Virtualization for High Performance Computing (Association for Computing Machinery, New York, NY, 2009), pp. 17–24.

[23] N. P. Jouppi et al., "In-Datacenter Performance Analysis of a Tensor Processing Unit," in Proceedings - International Symposium on Computer Architecture, 2017, vol. Part F1286, pp. 1–12.

[24] P. Kennedy, "Graphcore", https://www.servethehome.com/hands-on-with-a-graphcorec2- ipu-pcie-card-at-dell-tech-world/.

[25] "Syntiant", http://www.tinymlsummit.org/syntiant_7-25_meetup.pdf.

[26] S. Moore, See https://spectrum.ieee.org/tech-talk/semiconductors/processors/ first-programmable-memristor-computer for "MemryX" (last accessed October 15, 2019).

[27] "Mythic", https://www.mythic-ai.com/technology/.

[28] See https://www.mythic-ai.com/technology/ for Mythic.

[29] A. Makarov, V. Sverdlov, and S. Selberherr, "Emerging memory technologies: Trends, challenges, and modeling methods," Microelectron. Reliab. 52, 628–634 (2012).

[30] A. Chen, "A review of emerging non-volatile memory (NVM) technologies and applications," Solid-State Electron. 125, 25–38 (2016).

[31] C. Sung, H. Hwang, and I. K. Yoo, "Perspective: A review on memristive hardware for neuromorphic computation," J. Appl. Phys. 124(15), 151903 (2018).

[32] E. Kadric, D. Lakata, and A. Dehon, "Impact of parallelism and memory architecture on FPGA communication energy," ACM Trans. Reconfigurable Technol. Syst. 9(4), 30 (2016).

[33] M. Horowitz, "1.1 Computing's energy problem (and what we can do about it)," in 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC) (IEEE, 2014), pp. 10–14.

[34] M. A. Nahmias, T. Ferreira de Lima, A. N. Tait, H.-T. Peng, B. J. Shastri, and P. R. Prucnal, "Photonic multiply-accumulate operations for neural networks," IEEE J. Sel. Top. Quantum Electron. 26(1), 7701518 (2019).

[35] Neuromorphic Photonics, edited by P. R. Prucnal and B. J. Shastri (CRC Press, 2017).

[36] W. Bogaerts, D. Pérez, J. Capmany, D. A. B. Miller, J. Poon, D. Englund, F. Morichetti, and A. Melloni, "Programmable photonic circuits," Nature 586, 207–216 (2020).

[37] M. Smit et al., "An introduction to InP-based generic integrated technology," IOP Semicond. Sci. Technol. 29(8), 083001 (2014).

[38] C. Huang et al., Demonstration of Photonic Neural Network for Fiber Nonlinearity Compensation in Long-Haul Transmission Systems (OFC, San Diego, 2020), Postdeadline Th4C.6.

[39] Y. Shen et al., "Deep learning with coherent nanophotonic circuits," Nat. Photonics 11, 441–446 (2017).

[40] A. N. Tait et al., "Silicon photonic modulator neuron," Phys. Rev. Appl. 11(6), 064043 (2019).

[41] See https://lightmatter.co/ for Lightmatter.

[42] See https://www.lightelligence.ai/ for Lightelligence.

[43] See https://luminous.co/ for Luminous.

[44] See https://www.lumiphase.com/ for Lumiphase.

[45] See http://www.Groq.com for Groq.

[46] See https://www.anandtech.com/show/12815/cambricon-makers-of-huaweiskirinnpu- ip-build-a-big-ai-chip-and-pcie-card for Cambricon.

[47] See https://www.gyrfalcontech.ai/solutions/2803s/ for Gyrfalcon.

[48] G. Sarantoglou, M. Skontranis, and C. Mesaritakis, "All optical integrate and fire neuromorphic node based on single section, quantum dot laser," IEEE J. Sel. Top. Quantum Electron. 26, 1900310 (2019).

[49] J. Robertson, E. Wade, Y. Kopp, J. Bueno, and A. Hurtado, "Towards neuromorphic photonic networks of ultrafast spiking laser neurons," IEEE J. Sel. Top. Quantum Electron. 26, 7700715 (2019).

[50] S. Xiang, Z. Ren, Z. Song, Y. Zhang, X. Guo, G. Han, and Y. Hao, "Computing primitive of fully VCSEL-based All-optical spiking neural network

for supervised learning and pattern classification," IEEE Trans. Neural Netw. Learn. Syst. 1–12 (2020).

[51] J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, and W. H. P. Pernice, "All-optical spiking neurosynaptic networks with self-learning capabilities," Nature 569(7755), 208–214 (2019).

[52] R. Hamerly, L. Bernstein, A. Sludds, M. Soljačić, and D. Englund, "Large-scale optical neural networks based on photoelectric multiplication," Phys. Rev. X 9(2), 021032 (2019).

[53] L. Bernstein, A. Sludds, R. Hamerly, V. Sze, J. Emer, and D. Englund, "Freely scalable and reconfigurable optical hardware for deep learning," arXiv:2006.13926.

[54] H. Chaoran, S. Bilodeau, T. Ferreira de Lima, A. N. Tait, P. Y. Ma, E. C. Blow, A. Jha, H.-T. Peng, B. J. Shastri, and P. R. Prucnal, "Demonstration of scalable microring weight bank control for large-scale photonic integrated circuits," APL Photonics 5(4), 040803 (2020).

[55] G. Mourgias-Alexandris et al., "Neuromorphic photonics with coherent linear neurons using dual-IQ modulation cells," J. Lightwave Technol. 38(4), 811–819 (2020).

[56] G. Mourgias-Alexandris, A. Tsakyridis, N. Passalis, A. Tefas, K. Vyrsokinos, and N. Pleros, "An all-optical neuron with sigmoid activation function," Opt. Express 27, 9620–9630 (2019).

[57] L. Gelens et al., "Exploring multistability in semiconductor ring lasers: Theory and experiment," Phys. Rev. Lett., vol. 102, no. 19, pp. 1–4, 2009.

[58] H.-T. Peng, M. A. Nahmias, T. F. de Lima, A. N. Tait, and B. J. Shastri, "Neuromorphic Photonic Integrated Circuits," IEEE J. Sel. Top. Quantum Electron., vol. 24, no. 6, pp. 1–15, Nov. 2018.

[59] R. Stabile, A. Rohit, and K. A. Williams, "Dynamic multi-path WDM routing in a monolithically integrated 8 × 8 cross-connect," Opt. Express, vol. 22, no. 1, p. 435, 2014.

[60] K. Vandoorne, J. Dambre, D. Verstraeten, B. Schrauwen, and P. Bienstman, "Parallel reservoir computing using optical amplifiers," IEEE Trans. Neural Networks, vol. 22, no. 9, pp. 1469–1481, 2011.

[61] A. N. Tait et al., “Microring Weight Banks,” IEEE J. Sel. Top. Quantum Electron., vol. 22, no. 6, pp. 312–325, Nov. 2016.

[62] J. Han, J. Pei, and M. Kamber, Data mining: concepts and techniques. Elsevier, 2011.

[63] A. N. Tait, J. Chang, B. J. Shastri, M. A. Nahmias, and P. R. Prucnal, “Demonstration of WDM weighted addition for principal component analysis,” Opt. Express, vol. 23, no. 10, p. 12758, May 2015.

[64] B. Shi, N. Calabretta, D. Bunandar, D. Englund, and R. Stabile, “WDM Weighted Sum in an 8x8 SOA-Based InP Cross-Connect for Photonic Deep Neural Networks,” in 2018 Photonics in Switching and Computing (PSC), 2018, pp. 1–3.

[65] B. Shi, N. Calabretta, and R. Stabile, “SOA-Based Photonic Integrated Deep Neural Networks for Image Classification,” in Conference on Lasers and Electro-Optics, 2019, p. SF1N.5.

[66] R. A. Fisher, “The Use of Multipe Measurements in Taxonomic Problems,” Ann. Eugen., vol. 7, no. 2, pp. 179–188, Sep. 1936.

[67] M. Abadi et al., “TensorFlow: A System for Large-Scale Machine Learning,” in 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI ’16), 2016, pp. 265–283.

[68] A. N. Tait, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, “Broadcast and Weight: An Integrated Network For Scalable Photonic Spike Processing,” J. Light. Technol., vol. 32, no. 21, pp. 4029–4041, Nov. 2014.

[69] B. Shi, N. Calabretta, and R. Stabile, “SOA-Based Photonic Integrated Deep Neural Networks for Image Classification,” in 2019 Conference on Lasers and Electro-Optics, paper SF1N.5., 2019.

[70] Lowery, A.J. Amplified Spontaneous Emission in Semiconductor Laser Amplifiers. Validity of the Transmission-Line Laser Model. IEE proceedings. Part J. Optoelectron. 137, 241–247, 1990.

[71] M. Usami, M. Tsurusawa, and Y. Matsushima, “Mechanism for reducing recovery time of optical nonlinearity in semiconductor laser amplifier,” Appl. Phys. Lett., vol. 72, no. 21, p. 2657, May 1998.

[72] A. Mecozzi, S. Scotti, A. D'nOttavi, E. Iannone, and P. Spano, "Four-Wave Mixing in Traveling-Wave Semiconductor Amplifiers," IEEE J. Quantum Electron., vol. 31, no. 4, pp. 689–699, 1995.

[73] M. J. Connelly, Semiconductor Optical Amplifiers. Boston: Kluwer Academic Publishers, 2002.

[74] A. Bogoni and L. Potì, "Effective channel allocation to reduce inband FWM crosstalk in DWDM transmission systems," IEEE J. Sel. Top. Quantum Electron., vol. 10, no. 2, pp. 387–392, 2004.

[75] Y. Guo, B. Aazhang, and J. F. Young, "Wavelength encoding to reduce four-wave mixing crosstalk in multi-wavelength channels," Conf. Proc. - Lasers Electro-Optics Soc. Annu. Meet., vol. 2, pp. 230–231, 1997.

[76] D. D'Agostino, D. Lenstra, H. P. M. M. Ambrosius, and M. K. Smit, "Widely tunable Coupled Cavity Laser based on a Michelson Interferometer with doubled Free Spectral Range," in Optical Fiber Communication Conference, 2015, p. M2D.4.

[77] B. Shi, K. Prifti, E. Magalhães, N. Calabretta, and R. Stabile, "Lossless Monolithically Integrated Photonic InP Neuron for All-Optical Computation," in Optical Fiber Communication Conference (OFC) 2020, p. W2A.12.

[78] B. Shi, N. Calabretta, and R. Stabile, "First Demonstration of a Two-Layer All-Optical Neural Network by Using Photonic Integrated Chips and SOAs," in 45th European Conference on Optical Communication (ECOC 2019), 2019, pp. 398.

[79] Bin Shi, N. Calabretta, and R. Stabile, "Multi-Wavelength, Multi-Level Inputs for an All-Optical SOA-Based Neuron", in Conference on Lasers and Electro-Optics (CLEO) 2021, p. SM1B.4.

[80] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proc. IEEE, vol. 86, no. 11, pp. 2278–2323, 1998.

[81] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 2015, pp. 1–15.

[82] A. M. de Melo and K. Petermann, “On the amplified spontaneous emission noise modeling of semiconductor optical amplifiers,” Opt. Commun., vol. 281, no. 18, pp. 4598–4605, 2008.

[83] J. Choquette, W. Gandhi, O. Giroux, N. Stam, and R. Krashinsky, “NVIDIA A100 Tensor Core GPU: Performance and Innovation,” IEEE Micro, vol. 41, no. 2, pp. 29–35, 2021.

[84] H. Isono, “Latest standardization trend for high-speed optical transceivers with a view of beyond tera era,” in Metro and Data Center Optical Networks and Short-Reach Links III, vol. 1130808, no. January 2020, p. 7, 2020.

[85] K. Hosseini et al., “8 Tbps Co-Packaged FPGA and Silicon Photonics Optical IO,” Opt. Fiber Commun. Conf. 2021, p. Th4A.2, Jun. 2021.

[86] C. Zhang, P. Li, G. Sun, Y. Guan, B. Xiao, and J. Cong, “Optimizing FPGA-based accelerator design for deep convolutional neural networks,” in ACM/SIGDA International Symposium on FPGA, pp. 161–170, 2015.

[87] M. Erett et al., “A 0.5-16.3Gbps multi-standard serial transceiver with 219mW/channel in 16nm FinFET,” in European Solid-State Circuits Conference, pp. 297–300, 2016.

[88] A. Amara, F. Amiel, and T. Ea, “FPGA vs. ASIC for low power applications,” Microelectronics J., vol. 37, no. 8, pp. 669–677, Aug. 2006.

[89] S. Tanaka, S.-H. Jeong, S. Sekiguchi, T. Kurahashi, Y. Tanaka, and K. Morito, “High-output-power, single-wavelength silicon hybrid laser using precise flip-chip bonding technology,” Opt. Express, vol. 20, no. 27, p. 28057, 2012.

[90] X. Zheng et al., “A high-speed, tunable silicon photonic ring modulator integrated with ultra-efficient active wavelength control,” Opt. Express, vol. 22, no. 10, p. 12628, 2014.

[91] E. Swindlehurst et al., “An 8-bit 10-GHz 21-mW Time-Interleaved SAR ADC with Grouped DAC Capacitors and Dual-Path Bootstrapped Switch,” IEEE Solid-State Circuits Lett., vol. 2, no. 9, pp. 83–86, 2019.

[92] F. Y. Liu et al., “10-Gbps, 5.3-mW Optical Transmitter and Receiver Circuits in 40-nm CMOS,” IEEE J. Solid-State Circuits, vol. 47, no. 9, pp. 2049–2067, Sep. 2012.

[93] M. Sugawara, T. Akiyama, N. Hatori, Y. Nakata, H. Ebe, and H. Ishikawa, “Quantum-dot semiconductor optical amplifiers for high-bit-rate signal processing up to 160 Gb s-1 and a new scheme of 3R regenerators,” Meas. Sci. Technol., vol. 13, no. 11, pp. 1683–1691, 2002.

[94] H. Sun, Q. Wang, H. Dong, G. Zhu, N. K. Dutta, and J. Jaques, “Gain dynamics and saturation property of a semiconductor optical amplifier with a carrier reservoir,” IEEE Photonics Technol. Lett., vol. 18, no. 1, pp. 196–198, 2006.

[95] B. Shi, N.Calabretta. and R. Stabile, InP photonic integrated multi-layer neural networks: Architecture and performance analysis, APL Photonics 7 10801, 2022.

[96] K. Obermann, I. Koltchanov, K. Petermann, S. Diez, R. Ludwig and H. G. Weber, Noise analysis of frequency converters utilizing semiconductor-laser amplifiers IEEE J. Quantum Electron. 33 81–8, 1997.

[97] D. A. O. Davies, Small-signal analysis of wavelength conversion in semiconductor laser amplifiers via gain saturation IEEE Photonics Technol. Lett. 7 617–9, 1995.

[98] A. Mecozzi, Small-signal theory of wavelength converters based on cross-gain modulation in semiconductor optical amplifiers IEEE Photonics Technol. Lett. 8 1471–3, 1996.

[99] K. Obermann, I. Koltchanov, K. Petermann, C. Schmidt, S. Diez and H. G. Weber, Noise characteristics of semiconductor-optical amplifiers used for wavelength conversion via cross-gain and cross-phase modulation IEEE Photonics Technol. Lett. 9 312–4, 1997.

[100] K. Obermann, S. Kindt, D. Breuer and K. Petermann, Performance analysis of wavelength converters based on cross-gain modulation in semiconductor-optical amplifiers J. Light. Technol. 16 78–85, 1998.

[101] N. A. Olsson, Lightwave Systems With Optical Amplifiers J. Light. Technol. 7 1071–82, 1989.

[102] M. Miscuglio et al., “Massively parallel amplitude-only Fourier neural network,” Optica, vol. 7, no. 12, p. 1812, Dec. 2020.

[103] B. Rahmani, D. Loterie, G. Konstantinou, D. Psaltis, and C. Moser, “Multimode optical fiber transmission with a deep learning network,” Light Sci. Appl., vol. 7, no. 1, Dec. 2018.

[104] U. Teğin, M. Yıldırım, İ. Oğuz, C. Moser, and D. Psaltis, “Scalable optical learning operator,” Nat. Comput. Sci., vol. 1, no. 8, pp. 542–549, 2021.

[105] V. Bangari et al., “Digital Electronics and Analog Photonics for Convolutional Neural Networks (DEAP-CNNs),” IEEE J. Sel. Top. Quantum Electron., vol. 26, no. 1, 2020.

[106] C. Wu, H. Yu, S. Lee, R. Peng, I. Takeuchi, and M. Li, “Programmable phase-change metasurfaces on waveguides for multimode photonic convolutional neural network,” Nat. Commun., vol. 12, no. 1, p. 96, Dec. 2021.

[107] A. Mehrabian, M. Miscuglio, Y. Alkabani, V. J. Sorger, and T. El-Ghazawi, “A Winograd-Based Integrated Photonics Accelerator for Convolutional Neural Networks,” IEEE J. Sel. Top. Quantum Electron., vol. 26, no. 1, 2020.

[108] J. Feldmann et al., “Parallel convolutional processing using an integrated photonic tensor core,” Nature, vol. 589, no. 7840, pp. 52–58, 2021.

[109] E.W.R. Schultz, J.V. de Nijs, B. Shi, R. Stabile, “Optical 4F Correlator for Acceleration of Convolutional Neural Networks”, 25th Annual Symposium of the IEEE Photonics Society Benelux Chapter, 2021.

[110] E. Cottle, F. Michel, J. Wilson, N. New, and I. Kundu, “Optical Convolutional Neural Networks – Combining Silicon Photonics and Fourier Optics for Computer Vision,” 2020.

[111] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. K¨opf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “PyTorch: An imperative style, high-performance deep learning library,” in Advances in Neural Information Processing Systems, vol. 32. Neural information processing systems foundation, 12 2019.

[112] H. Pratt, B. Williams, F. Coenen, and Y. Zheng, “FCNN: Fourier Convolutional Neural Networks,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10534 LNAI, 2017, pp. 786–798.

[113] X. Xu et al., “11 TOPS photonic convolutional accelerator for optical neural networks,” Nature, vol. 589, no. 7840, pp. 44–51, 2021.

[114] Y. Jiao et al., “InP membrane integrated photonics research,” Semicond. Sci. Technol., vol. 36, no. 1, 2020.

[115] W. R. Clements, P. C. Humphreys, B. J. Metcalf, W. S. Kolthammer, and I. A. Walmsley, “Optimal design for universal multiport interferometers,” Optica, vol. 3, no. 12, 2016.

[116] S. Pai, B. Bartlett, O. Solgaard, and D. A. B. Miller, “Matrix Optimization on Universal Unitary Photonic Devices,” Phys. Rev. Appl., vol. 11, no. 6, p. 1, 2019.

[117] C. Huang et al., “Programmable Silicon Photonic Optical Thresholder,” IEEE Photonics Technol. Lett., vol. 31, no. 22, pp. 1834–1837, 2019.

[118] Y. Jiao et al., “Indium Phosphide Membrane Nanophotonic Integrated Circuits on Silicon,” Phys. Status Solidi Appl. Mater. Sci., vol. 217, no. 3, pp. 1–12, 2020.

[119] O. Fagbohungbe and L. Qian, “Benchmarking Inference Performance of Deep Learning Models on Analog Devices,” pp. 1–9, 2020.

[120] R. Ryf et al., “Wavelength blocking filter with flexible data rates and channel spacing,” J. Light. Technol., vol. 23, no. 1, pp. 54–61, Jan. 2005.

[121] K. Prifti, X. Xue, N. Tessema, R. Stabile, and N. Calabretta, “Lossless Photonic Integrated Add-Drop Switch Node for Metro-Access Networks,” IEEE Photonics Technol. Lett., vol. 32, no. 7, pp. 387–390, 2020.

[122] M. Miscuglio et al., “Artificial Synapse with Mnemonic Functionality using GSST-based Photonic Integrated Memory,” in 2020 International Applied Computational Electromagnetics Society Symposium (ACES), 2020, pp. 1–3.

[123] N. C. Harris et al., “Efficient, compact and low loss thermo-optic phase shifter in silicon,” Opt. Express 22, 10487–10493 (2014).

[124] Z. Lu, K. Murray, H. Jayatilleka, and L. Chrostowski, “Michelson interferometer thermo-optic switch on SOI with a 50-µW power consumption,” in 2016 IEEE Photonics Conference (IPC), Waikoloa, HI (IEEE, 2016), pp. 107–110.

[125] A. Totovic et al., “Femtojoule per MAC neuromorphic photonics: An energy and technology roadmap,” IEEE J. Sel. Top. Quantum Electron. 26(5), 8800115 (2020).

[126] C. Wang, M. Zhang, M. Yu, R. Zhu, H. Hu, and M. Loncar, Nat. Commun. 10, 978 (2019).

[127] I. Bar-Joseph, C. Klingshirn, D. A. B. Miller, D. S. Chemla, U. Koren, and B. I. Miller, Appl. Phys. Lett. 50, 1010 (1987).

[128] Y. Kuo, Y. K. Lee, Y. Ge, S. Ren, J. E. Roth, T. I. Kamins, D. A. B. Miller, and J. S. Harris, IEEE J. Sel. Top. Quantum Electron. 12, 1503 (2006).

[129] M. R. Billah, M. Blaicher, T. Hoose, P.-I. Dietrich, P. Marin-Palomo, N. Lindenmann, A. Nesic, A. Hofmann, U. Troppenz, M. Moehrle, S. Randel, W. Freude, and C. Koos, Optica 5, 876 (2018).

[130] R. Amin, R. Maiti, C. Carfano, Z. Ma, M. H. Tahersima, Y. Lilach, D. Ratnayake, H. Dalir, and V. J. Sorger, APL Photonics 3, 126104 (2018).

[131] C. Wang, M. Zhang, X. Chen, M. Bertrand, A. Shams-Ansari, S. Chandrasekhar, P. Winzer, and M. Lončar, Nature 562, 101 (2018).

[132] C. Haffner, D. Chelladurai, Y. Fedoryshyn, A. Josten, B. Baeuerle, W. Heni, T. Watanabe, T. Cui, B. Cheng, S. Saha, D. L. Elder, L. R. Dalton, A. Boltasseva, V. M. Shalaev, N. Kinsey, and J. Leuthold, Nature 556, 483 (2018).

[133] A. Grigorenko, M. Polini, and K. Novoselov, “Graphene plasmonics,” Nat. Photonics 6, 749–758 (2012).

[134] S. Ohno, Q. Li, N. Sekine, J. Fujikata, M. Noguchi, S. Takahashi, K. Toprasertpong, S. Takagi, and M. Takenaka, “Taper-less III-V/Si hybrid MOS optical phase shifter using ultrathin InP membrane,” in Optical Fiber Communication Conference (OFC) 2020, OSA Technical Digest (Optical Society of America, 2020), p. M2B.6.

[135] N. Youngblood and M. Li, “Integration of 2D materials on a silicon photonics platform for optoelectronics applications,” Nanophotonics 6(6), 1205–1218 (2016).

[136] R. Amin, R. Maiti, Y. Gui, C. Suer, M. Miscuglio, E. Heidari, R. T. Chen, H. Dalir, and V. J. Sorger, “Sub-wavelength GHz-fast broadband ITO

Mach– Zehnder modulator on silicon photonics," Optica 7, 333–335 (2020).

[137] K. J. Laboy-Juárez, S. Ahn, and D. E. Feldman, "A normalized template matching method for improving spike detection in extracellular voltage recordings," Sci. Rep. 9, 12087 (2019).

[138] K. Kato, M. Kuwahara, H. Kawashima, T. Tsuruoka, and H. Tsuda, "Current-driven phase-change optical gate switch using indium-tin-oxide heater," Appl. Phys. Express 10(7), 072201 (2017).

[139] H. Zhang et al., "Miniature multilevel optical memristive switch using phase change material," ACS Photonics 6(9), 2205–2212 (2019).

[140] J. Zheng et al., "Nonvolatile electrically reconfigurable integrated photonic switch enabled by a silicon PIN diode heater," Adv. Mater. 32, 2001218 (2020).

[141] M. Delaney, I. Zeimpekis, D. Lawson, D. W. Hewak, and O. L. Muskens, "A new family of ultralow loss reversible phase-change materials for photonic integrated circuits: Sb2S3 and Sb2Se3," Adv. Funct. Mater. 30, 2002447 (2020).

[142] A. N. Tait et al., "Multi-channel microring weight bank control for reconfigurable analog photonic networks," in 2016 IEEE Optical Interconnects Conference (OI) (IEEE, 2016), pp. 104–105.

[143] D. Dagostino, G. Carnicella, C. Ciminelli et al., "Low loss waveguides for standardized InP integration processes," in Proceedings of the 18th Annual Symposium of the IEEE Photonics Society Benelux Chapter (Technische Universiteit Eindhoven, 2013).

[144] T. D. Bucio et al., "Silicon nitride photonics for the near-infrared," IEEE J. Sel. Top. Quantum Electron. 26(2), 8200613 (2019).

[145] T. D. Bucio, A. Z. Khokhar, G. Z. Mashanovich, and F. Y. Gardes, "N-rich silicon nitride angled MMI for coarse wavelength division (de)multiplexing in the O-band," Opt. Lett. 43(6), 1251 (2018).

[146] J. Chiles, S. M. Buckley, S. W. Nam, R. P. Mirin, and J. M. Shainline, "Design, fabrication, and metrology of 10 × 100 multi-planar integrated photonic routing manifolds for neural networks," APL Photonics 3, 106101 (2018).

[147] T. F. de Lima, A. N. Tait, H. Saeidi, M. A. Nahmias, H. Peng, S. Abbaslou, B. J. Shastri, and P. R. Prucnal, “Noise analysis of photonic modulator neurons,” IEEE J. Sel. Top. Quantum Electron. 26(1), 7600109 (2020).

[148] R. Amin et al., “ITO-based electro-absorption modulator for photonic neural activation function,” APL Mater. 7(8), 081112 (2019).

[149] M. M. P. Fard, I. A. D. Williamson, M. Edwards, K. Liu, S. Pai, B. Bartlett, M. Minkov, T. W. Hughes, S. Fan, and T.-A. Nguyen, “Experimental realization of arbitrary activation functions for optical neural networks,” Opt. Express 28, 12138–12148 (2020).

[150] M. Miscuglio, A. Mehrabian, Z. Hu, S. I. Azzam, J. George, A. V. Kildishev, M. Pelton, and V. J. Sorger, “All-optical nonlinear activation function for photonic neural networks [invited],” Opt. Mater. Express 8, 3851–3863 (2018).

[151] K. Nozaki et al., “Ultracompact O-E-O converter based on fF-capacitance nanophotonic integration,” in 2018 Conference on Lasers and Electro-Optics (CLEO), San Jose, CA (IEEE, 2018), pp. 1–2.

[152] M. Miscuglio et al., “Artificial synapse with mnemonic functionality using GSST-based photonic integrated memory,” arXiv:1912.02221 (2019).

[153] D. A. Bekele et al., “Signal reshaping and noise suppression using photonic crystal Fano structures,” Opt. Express 26, 19596 (2018).

[154] A. Jha, C. Huang, and P. R. Prucnal, “Reconfigurable all-optical nonlinear activation functions for neuromorphic photonics,” Opt. Lett. 45(17), 4819–4822 (2020).

[155] A. Jha, C. Huang, T. Ferreira de Lima, and P. R. Prucnal, “High-speed all-optical thresholding via carrier lifetime tunability,” Opt. Lett. 45(8), 2287–2290 (2020).

[156] M. Ono, M. Hata, M. Tsunekawa et al., “Ultrafast and energy-efficient all-optical switching with graphene-loaded deep-subwavelength plasmonic waveguides,” Nat. Photonics 14, 37–43 (2020).

[157] A. Bazin, P. Monnier, X. Lafosse, G. Beaudoin, R. Braive, I. Sagnes, R. Raj, and F. Raineri, “Thermal management in hybrid InP/silicon photonic crystal nanobeam laser,” Opt. Express 22, 10570–10578 (2014).

[158] G. Crosnier, D. Sanchez, S. Bouchoule, P. Monnier, G. Beaudoin, I. Sagnes, R. Raj, and F. Raineri, “Hybrid indium phosphide-on-silicon nanolaser diode,” Nat. Photonics 11(5), 297 (2017).

[159] K. Lengle et al., “Modulation contrast optimization for wavelength conversion of a 20 Gbit/s data signal in hybrid InP/SOI photonic crystal nanocavity,” Opt. Lett. 39(8), 2298 (2014).

[160] K. Nozaki et al., “Photonic-crystal nano-photodetector with ultrasmall capacitance for on-chip light-to-voltage conversion without an amplifier,” Optica 3, 483–492 (2016).

[161] T. Alexoudi et al., “III–V-on-Si photonic crystal nanocavity laser technology for optical static random access memories,” IEEE J. Sel. Top. Quantum Electron. 22(6), 295–304 (2016).

Publication List

Journal paper

1. B. Shi, N. Calabretta, and R. Stabile, “Deep Neural Network Through an InP SOA-Based Photonic Integrated Cross-Connect,” IEEE J. Sel. Top. Quantum Electron., vol. 26, no. 1, p.7701111, Jan. 2020.
2. B. Shi, N. Calabretta, and R. Stabile, “Numerical Simulation of an InP Photonic Integrated Cross-Connect for Deep Neural Networks on Chip **[Invited]**,” Appl. Sci., vol. 10, no. 2, p. 474, Jan. 2020.
3. B. Shi, N. Calabretta, and R. Stabile, “InP photonic integrated multi-layer neural networks: Architecture and performance analysis **[Invited]**,” APL Photonics, vol. 7, no. 1, p. 10801, 2022.
4. B. Shi, N. Calabretta, and R. Stabile, “Emulation and Modelling of SOA-based All-Optical Photonic Integrated Deep Neural Network with Arbitrary Depth **[Invited]**,” submitted to IOP Neuromorphic Computing Engineering, 2022.
5. B. Shi, N. Calabretta, R. Stabile, “Photonic Convolutional Neural Networks though SOA-based Cross-connect **[Invited]**”, for the IEEE J. Sel. Top. Quantum Electron. Special issue on Optical Computing, in preparation.
6. R. Stabile, G. Dabos, C. Vagionas, B. Shi, N. Calabretta, and N. Pleros, “Neuromorphic photonics: 2D or not 2D?” J. Appl. Phys., vol. 129, no. 20, p. 200901, 2021.
7. D. W Feyisa, B. Shi, R. Kraemer, K.A William, N. Calabretta and R. Stabile. Compact Lossless 8×8 SOA-Based Optical WDM Space Switch in Generic InP Technology, submitted to Journal of Lightwave Technology, 2022.

Conference paper

1. B. Shi, N. Calabretta, D. Bunandar, D. Englund, and R. Stabile, “WDM Weighted Sum in an 8x8 SOA-Based InP Cross-Connect for Photonic Deep Neural Networks,” in 2018 Photonics in Switching and Computing (PSC),

2018, pp. 1–3.

2. B. Shi, N. Calabretta, and R. Stabile, “Integrated semiconductor optical amplifiers based photonic cross-connect for deep neural networks,” in Conference in Cognitive Computing: Merging Concepts with Hardware, 2018.
3. B. Shi, N. Calabretta, and R. Stabile. "Continuous weight tuning for WDM-based neuron addition in an SOA-based InP cross-connect." 23rd Annual Symposium of the IEEE Photonics Society Benelux Chapter, 2018.
4. B. Shi, N. Calabretta, and R. Stabile, “Two-layer all-optical deep neural network with photonic integrated weighted addition,” European Materials Research Society fall meeting, 2019.
5. B. Shi, N. Calabretta, and R. Stabile, "SOA-Based Photonic Integrated Deep Neural Networks for Image Classification," in Conference on Lasers and Electro-Optics, paper SF1N.5., 2019.
6. B. Shi, N. Calabretta, and R. Stabile, "InP Photonic Circuit for Deep Neural Networks," in OSA Advanced Photonics Congress (AP) 2019 (IPR, Networks, NOMA, SPPCom, PVLED), paper IW2A.3., 2019.
7. B. Shi, N. Calabretta, and R. Stabile, “Image Classification with a 3-Layer SOA-Based Photonic Integrated Neural Network,” in OECC/PSC, p. MG1-1, 2019.
8. B. Shi, N. Calabretta, and R. Stabile. "Error analysis of a 3-Layer SOA-based photonic deep neural network for image classification." 24th annual Symposium of the IEEE Photonics Benelux Chapter. 2019.
9. B. Shi, N. Calabretta, and R. Stabile, “First Demonstration of a Two-Layer All-Optical Neural Network by Using Photonic Integrated Chips and SOAs,” in 45th European Conference on Optical Communication (ECOC 2019), 2019.
10. B. Shi, N. Calabretta, and R. Stabile, “Matrix Multiplication Unit Scalability Investigation for InP SOA-based Photonic Deep Neural Network **[Invited]**,” IEEE Photonics Soc. Summer Top. Meet. Ser., 2020.
11. B. Shi, K. Prifti, E. Magalhães, N. Calabretta, and R. Stabile, “Lossless Monolithically Integrated Photonic InP Neuron for All-Optical Computation,” in Optical Fiber Communication Conference (OFC), p. W2A.12, 2020.
12. B. Shi, B. Pan, N. Calabretta, and R. Stabile, “Multi-Wavelength, Multi-

Level Inputs for an All-Optical SOA-Based Neuron," Conf. Lasers Electro-Optics, p. SM1B.4, 2021.

13. B. Shi, B. Pan, N. Calabretta, and R. Stabile, "Noise and Scalability Investigation of SOA-based All-optical Photonic Deep Neural Network," in Asia Commun. Photonics Conf., p. T1E.4, 2021.

14. B. Shi, B. Pan, N. Calabretta, and R. Stabile, "Noise analysis of SOA-based All-optical Photonic Deep Neural Network with WDM input," in Photonics Switch. Comput. Conf., p. Tu3B.4, 2021.

15. B. Shi, B. Pan, N. Calabretta, and R. Stabile, "Scalability Analysis of the SOA-based All-optical DeepNeural Network", 25th annual Symposium of the IEEE Photonics Benelux Chapter. 2021.

16. B. Shi, N. Calabretta, and R. Stabile, 'SOA-based All-optical Photonic Integrated Deep Neural Network with Stable Output Noise', accepted to 48th European Conference on Optical Communication (ECOC 2022), 2022.

17. E.W.R. Schultz, J.V. de Nijs, B. Shi, R. Stabile, "Optical 4F Correlator for Acceleration of Convolutional Neural Networks", 25th Annual Symposium of the IEEE Photonics Society Benelux Chapter, 2021.

18. R. Stabile, N. Calabretta, and B. Shi, "Large-Scale Photonic Integrated Cross-Connects for Optical Communication and Computation **[Invited]**," Opt. Fiber Commun. Conf., p. Th3B.1, 2020.

19. D. W. Feyisa, B. Shi, B. Smalbrugge, K. A. Williams, and R. Stabile, "Ultra-compact 8 × 8 optical space switch with generic InP technology," Int. Conf. Transparent Opt. Networks, p. We.A5.6, 2020.

20. R. Stabile, N. Tessema, K. Prifti, D. W. Feyisa, B. Shi, and N. Calabretta, "Dense Photonic InP Integration for Modular Nodes in Next Generation Optical Networks [**Invited]**," in OSA Advanced Photonics Congress (AP) 2020 (IPR, NP, NOMA, Networks, PVLED, PSC, SPPCom, SOF), p. NeTh1B.1., 2020.

21. R. Stabile, N. Tessema, K. Prifti, D. W. Feyisa, B. Shi, and N. Calabretta, "Photonic integrated nodes for next-generation metro optical networks **[Invited]**," in Metro and Data Center Optical Networks and Short-Reach Links IV, vol. 11712, p.10, 2021.

22. K. Prifti, N. Tessema, B. Shi, A. R. Zali, S. Kleijn, L. Augustin, R. Stabile,

and N. Calabretta, “Evaluation of a 1×8 Photonic Integrated WDM Wavelength Selective Switch for Optical Data Center Networks,” in 26th Optoelectronics and Communications Conference (OECC), p. S4A.3. 2021.

23. D. W. Feyisa, B. Shi, B. Smalbrugge, K. A. Williams, N. Calabretta, and R. Stabile, “140 Gb/s WDM Data Routing in a Lossless Strictly Non-Blocking SOA-Based Photonic Integrated 8×8 Space Switch,” in Optical Fiber Communications Conference and Exhibition (OFC), p. W1C.1, 2021.

Patent

1. B. Shi, R. Stabile and N. Calabretta, “Integrated Multi-Wavelength Conversion based Noise Stable Optical Branching”, in preparation.

Acknowledgements

It has been a long journey to reach the end of my Ph.D. research, albeit it seems so short since I am still engaged in the exciting research. A poem from the 楚辭 *Chuci* collection can express my emotion explicitly at this moment:

‘朝發軔於蒼梧兮，夕余至乎縣圃。欲少留此靈瑣兮，日忽忽其將暮。
吾令羲和弭節兮，望崦嵫而勿迫。路曼曼其脩遠兮，吾將上下而求索。’

‘In the morning I started on my way from Ts'ang-wu; In the evening I came to the Garden of Paradise.
I wanted to stay a while in those fairy precincts, But the swift-moving sun was dipping to the west.
I order Hsi-ho (sun-Charioteer) to stay the sun-steed's gallop, To stand over Yen-tzu mountain not go in.
Long, long had been my road and far, far was the journey; I would go up and down to seek my heart's desire.’
(屈原 Qu Yuan, ~300 B.C., translated by David Hawkes, 1959)

Coincidently, Ts'ang-wu is the ancient name of my hometown, Wuzhou. In the imagination, the poet Qu Yuan started a long journey from Ts'ang-wu to the Garden of Paradise to seek his heart's desire and hope the time goes slower. Similarly, the marvel of neuromorphic photonics makes me enjoy the works in the ECO group and feel that time flies. I am fortunate to be supported by so many people to complete my PhD works. I would like to express my sincere gratitude to all these people without whom this dissertation would not be possible.

First of all, my deepest gratitude goes to my first promotor, dr. Ripalta (Patty) Stabile, for offering me the position to let me pursue my Ph.D. for exploring the cutting-edge field of Neuromorphic Photonics. I sincerely thank you for giving me the freedom to develop different ideas and for providing advices, encouragements and helps throughout my research. I am very grateful for the innumerous discussions we had on the projects and for your revisions on the papers that helps to shape the research story presented in this dissertation. Your profession in optical switches, optical integration, and optical computing has inspired and supported my studies in many degrees. And your instructions for building the photonic integrated circuit testing set up and the help on the very first alignment of the lensed fiber gave me precious experiences that helped a lot for further

experimental setup development. Moreover, your open-mindedness, kindness and patience will influence me as a person. I really appreciated it, especially when sometimes I sink too deep into the challenges.

I would also like to thank my co-promotor, dr. Nicola Calabretta, who is always enthusiastic in research and providing valuable suggestions. I always benefit from our discussions and debates. I am grateful for your contributions to the meetings and the revision of the papers. Your inspiration and encouragement on research novelty help me to sort out my ideas. Your expertise in semiconductor optical amplifier and optical switch helps me a lot in analysing the PIC performance. I would like to thank my co-promotor, prof. Ton Koonen, for the general guidance and supports in the ECO group and the supports in building my experimental setup. I am also grateful for your evaluation of my progress, giving constructive suggestions for improving my professional skills, as well as giving comments for improving my Ph.D. dissertation.

I would use this opportunity to thank all the committee members for reviewing and accessing my works in this dissertation. I especially thank prof. Jose Capmany (Universitat Politecnica de Valencia), prof. Lorenzo Pavesi (Università degli Studi di Trento), prof. Bert Jan Offrein (Universiteit Twente), and dr. Sander Stuijk, for their suggestions that help improve my dissertation.

I would like to express my gratitude to dr. Darius Bunandar and dr. Dirk Englund from Massachusetts Institute of Technology, who helped to kick off my research and provided valuable support at my first stage of Ph.D. I am also grateful to dr. George Dabos, dr. Chris Vagionas and dr. Nikos Pleros, from the Aristotle University of Thessaloniki, for fruitful discussions and contributions on the perspective of the Neuromorphic Photonic.

I am thankful to my colleagues who contributed to part of my research outputs, with whom my research is more enjoyable. I especially thank Kristif Prifti for the sharing of the components and discussions on simulation tools and the experiment on the all-optical neuron testing. And thank dr. Eduardo Magalhães for the help on SOA simulation as well as wavelength converter operations. I am especially thankful to dr. Bitao Pan for the discussion on FPGA control and many other experimental ideas. I enjoy the experiments we did with your FPGA-based setup and the coffee or lunch time for sharing views and interests. I wish all of you great success in your new positions. I would also thank my former students Eloy Schultz and Joris de Nijs, for the excellent works on the free-space convolutional neural network. And also thank Desalegn Wolde Feyisa, for the great works on optical switches and the discussions on the simulations. I enjoy a lot in the laboratory with all of you and I wish you great success of in next studies or careers.

I am very grateful to other members in the ECO group. I am especially thankful to dr. Eduward Tangdiongga, for the maintenance for the Arbitrary Waveform Generator and Digital Phosphor Oscilloscope, which are essential devices for all my experiments. Many thanks to dr. Federico Forni for the instructions on using the devices mentioned above. Also thanks to dr. Junyu (Tom) Song, for teaching me the operation of optical modulators and helping in the optical signal generation and detection. I would thank our technician Johan van Zantvoort for the help on the purchases of devices/components and on the operations of the testing setup. You are always helpful with laboratory issues. And many thanks to technician Frans Huijskens for the help in workshop staff and fiber/optic components. It was impressive that you have fussed a fiber with a precise length that synchronised the optical signal for one of my experiments. I am thankful to dr. Ketemaw Mekonnen for setting up and maintenance of the simulation server.

I would like to thank our secretary Ms. José Hakkens, for your warm assistance of variety administrative matters, answering my general questions and solving my problems. I can always receive practical suggestions from you. Many thanks to Ms. Ginny Toes for helping the arrangement of my last phase of PhD. I am also thankful to Ms. Yvonne van Bokhoven, Ms. Tanja van Waterschoot and other colleagues from the HR back office and staff immigration team: they have help me a lot regarding the contract and visa issues.

Many thanks to dr. Simone Cardarelli, who works in the same laboratory, for the helps on the first automatic testing programming for data acquisition. I am also lucky to witness your great works in fiber to photonic chip coupling. I still remember the joy moment at the first time that you show me the ‘dancing’ v-groove piezo actuators for fiber arrays alignment. I wish you and your spin-off, MicroAlign, a great future.

I would thank the colleagues who works in the other labs, and I am grateful for the inspirations from them. I am thankful to dr. Xuebing Zhang, dr. Mahir Mohammed, dr. Xuwei Xue, dr. Chao Li, and dr. Mingyang Zhao, for the inspirated discussions and helps on experimental setups. I would be grateful to dr. Netsanet Tessema for the discussions on design of photonic integrated circuits and the characterisation. I am thankful to dr. Haotian Zeng, dr. Sjoerd van der Heide, Gianluca Kosmella, Aliee Trinidad, Menno van den Hout, Ngoc Quan Pham, Catalina Stan, and Oumayma Bouchmal, for introductions on your systems/research and providing helpful information on the components/methods. Many thanks to Yu Wang for the enjoyable experiments and coffee time for sharing stories and sorrows, and also to Aref Rasoulzadehzali, Yuchen Song, Shaojuan (Jessie) Zhao, Rafael Kreamer, Liuyan Chen, Yu Lei, Xinda Yan, Shiyi Xia, Zhiyu Chen, Jiangrui (Watson) Deng, Mohammad Mukit, and Antonio Astorino, for the

discussions and sharing of the common devices and components as well as the inspirations, chats, and jokes at the breaks. I would thank my office mates, dr. Xiaotao Guo, dr. Yu Zhao, dr. Jianou Huang, dr. Tim van der Lee, Ruby Ospina, Elham Khani, Javier Santacruz, and Asterios Souftas for the good time spending in the office works. Many thanks to dr. Fulong Yan who is always willing to have a discussion on my research and provides suggestions. I wish all of you all the best in your works and lives.

I would thank the general supports from the members of the group. I am grateful to dr. Zizheng (Pang) Cao, for the interesting discussions on my projects and the encouragement on my works. Many thanks to dr. Chigo Okonkwo and dr. Oded Raz, who are always enthusiastic about research and education, for their kind suggestions and supports. I am also grateful to prof. Idelfonso Monroy, prof. Sonia de Groot, dr. George Exarchakos, dr. Henrie van den Boom, dr. Hugo de Waardt, dr. Simon Rommel, dr. Shihab Al-Daffaie, for keeping the active research communications within the group. Also thanks to all other former and recent ECO members, dr. Dimitrios Konstantino, dr. Fu Wang, dr. Bruno Cimoli, dr. John van Weerdenburg, dr. Aaron Albores-Mejia, dr. Marc Spiegelberg, dr. Maria Hermelo, dr. Tom Bradley, dr. Amado Benitez, Henrique Santana, Yuzhe Wang, Panagiotis Giannakopoulos, Hamid Hassani, Hui Liu, Gleb Nazarikov, Maira Sosa, Carina Correa, Marijn Rombouts, Carolina Amaral, Mehmet Temel, Jaap Verheggen, Mikolaj Wolny, and Xinran Zhao, for the pleasant encounters.

I would express my gratitude to the colleagues from PhI group. Especially I would thank dr. Yuqing Jiao for helping my design on IMOS and providing simulation tools. And also to Yi Wang, who helps me a lot in the photonic circuit design as well as on the fabrications. I enjoy the lunch meeting and coffee break with you. I wish you great success in your research. Many thanks to dr. Erwin Bente for the help on the current control at the beginning of my PhD. And thanks to dr. Jos van der Tol and prof. Martijn Heck for the interesting discussions on photonic integration and transfer printing. Many thanks to dr. Tianran Liu for the discussion on computer-generated holography and on the fiber arrays. I am grateful to dr. Weiming Yao and dr. Victor Calzadilla for the talks on neurophotonic photonics as well as their useful professional development suggestions. I am grateful to dr. Vadim Pogoretskiy, dr. Vikram Bhagavatula, Lukas Puts, Ekaterina Malysheva, Riu Ma, Joel Hazan, Wenjing Tian, Rachel Jones, Yihui Wei, and dr. Limeng Zhang, for sharing the research ideas and sometimes inspiring my works.

I am grateful to the Chinese community. I would thank dr. Teng Li for introducing me to all the Chinese colleagues. Many thanks to dr. Chenhui Li, Lu Huang, dr. Bin Chen, dr. Wang Miao, dr. Qiang Liu, dr. Xiong Deng, Pan Gao, and Ping Bai etc. Together with the friends in our group, they gave me a great

welcome in the Netherlands upon my arrival and had a good time with beers during my PhD. I am also grateful to my friends in China, Wenchao Ma, Tongyi Sun, and Zichun Wang. I enjoyed the great time with them when I was back to China.

Last but not least, I would express my thanks to my parents, my sisters, and my extended family for their love and supports throughout the years. My special thanks go to my darling partner Yanying (Jana) Wu, who suggested, encouraged, and accompanied me to go across the continent to Europe and pursue my heart's desire in research. 感谢家人们的支持让我顺利完成学业。

Bin Shi 石彬
Eindhoven, June 2022

Curriculum Vitae

Bin Shi was born on the 8th of March 1992, in Wuzhou, China. He received the B.S. degree in optical information science and technology from China University of Mining and Technology, Xuzhou, China, in 2014, with a bachelor thesis entitled, ‘Simulation of Self-collimated Beam Coupling between Two Photonic Crystals’. He received his M.S. degree in Lasers and Photonics from Ruhr University Bochum, Bochum, Germany, in 2017. The topic of his Master thesis was ‘Two-photon Laser Terahertz Emission Microscopy with an OSCAT system’.

In March 2018, he started working toward the Ph.D. degree in Electro-Optical Communication (ECO) group, Institute of Photonic Integration, Eindhoven University of Technology, Eindhoven, the Netherlands. His research interests include photonic neural network, photonic integrated circuits, and optical signal processing. In his research, a photonic deep neural network is demonstrated with SOA-based cross-connect for pattern classification. And an all-optical neuron is investigated with SOAs working in linear and nonlinear regimes. He served as a teaching assistant involved in two courses and several bachelor ending projects and master ending projects in Electrical Engineering department of Eindhoven University of Technology.

He has authored and co-authored more than 25 peer-reviewed journal and international conference papers, including four invited papers. He is an active student member of IEEE Photonics Society and Optical Society of America, serving as a reviewer in journals including Photonics Technology Letters, Journal of Selected Topics in Quantum Electronics, and Proceedings of the IEEE.

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY