Emulation and modelling of semiconductor optical amplifier-based all-optical photonic integrated deep neural network with arbitrary depth

NEUROMORPHIC
Computing and Engineering

PAPER

Emulation and modelling of semiconductor optical amplifier-based all-optical photonic integrated deep neural network with arbitrary depth

Bin Shi[1,2,*], Nicola Calabretta[1] and Ripalta Stabile[1,2]

[1] Eindhoven Hendrik Casimir Institute, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands
[2] Eindhoven AI Systems Institute, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands
[*] Author to whom any correspondence should be addressed.

E-mail: b.shi1@tue.nl

Keywords: neuromorphic photonics, photonic neural network, photonic integrated circuits

OPEN ACCESS

RECEIVED
18 March 2022

REVISED
23 June 2022

ACCEPTED FOR PUBLICATION
9 August 2022

PUBLISHED
2 September 2022

Abstract

We experimentally demonstrate the emulation of scaling of the semiconductor optical amplifier (SOA) based integrated all-optical neural network in terms of number of input channels and layer cascade, with chromatic input at the neuron and monochromatic output conversion, obtained by exploiting cross-gain-modulation effect. We propose a noise model for investigating the signal degradation on the signal processing after cascades of SOAs, and we validate it via experimental results. Both experiments and simulations claim that the all-optical neuron (AON), with wavelength conversion as non-linear function, is able to compress noise for noisy optical inputs. This suggests that the use of SOA-based AON with wavelength conversion may allow for building neural networks with arbitrary depth. In fact, an arbitrarily deep neural network, built out of seven-channel input AONs, is shown to guarantee an error minor than 0.1 when operating at input power levels of −20 dBm/channel and with a 6 dB input dynamic range. Then the simulations results, extended to an arbitrary number of input channels and layers, suggest that by cascading and interconnecting multiple of these monolithically integrated AONs, it is possible to build a neural network with 12-inputs/neuron 12 neurons/layer and arbitrary depth scaling, or an 18-inputs/neuron 18-neurons/layer for single layer implementation, to maintain an output error <0.1. Further improvement in height scalability can be obtained by optimizing the input power.

1. Introduction

The exponential increase of data generation demands more effective data processing, not only on the computation speed but also on the power efficiency. The mismatch between the growth of data volume and the computation power has led to an exploration of complementary computing structures, also called non-von Neumann architectures. Brain-inspired computation structures have been developed to enable effective information extraction, while Moore's law is approaching its end [1]. Parallel micro-processors like graphic processing units [2], tensor processing units [3] and neuromorphic electronics based on application-specific integrated circuits [4–6] have been reducing the power consumption down to the pJ per operation. However, the electrical processing of massive data requires unprecedented computation power which the scaling of the transistors cannot match, both in terms of computation capacity and power efficiency [7], because of the power dissipation spent to move the electrical signal as well as of the limited interconnection bandwidth. Neuromorphic photonics is emerging as one of the approaches to facilitate the high-speed computation for artificial neural networks, taking advantage of the advanced parallelism of light. Recently, Mach–Zehnder interferometer based optical interference unit [8], large-scale computation obtained via free-space optics with balanced detection schemes [9], and the in-phase and quadrature modulation based computation unit [10] have been proposed for realizing photonic neural networks, exploiting the coherent approaches. A wavelength division multiplexing (WDM) approach enables high bandwidth interconnection between neurons. The use of broadcast-and-weight using mirroring weighting banks in [11, 12] has enabled a straightforward optical

power weight addition implementation but the synaptic operation may not ensure a large enough dynamic range when moving to subsequent network layers. Moreover, the use of phase change material (PCM), together with integrated cross-bar matrix, interfaced with frequency comb laser [13], can provide highly efficient convolutional signal processing. Although the PCM-based processing unit only consumes power when writing the PCM, the detection, the nonlinear processing, the data regeneration and the connection from layer to layer are still lying on electric signals. However, the scalability of the optical neural networks is still an open question since most of the layer-to-layer interconnection relies on the electro-optical (E/O) conversions [8–13], which leads to costly and power-consuming electronics design or limits bandwidth. An all-optical approach for layer-to-layer interconnection can overcome this problem. The PCM-based neural network in [14] is proposed to enable all-optical interconnection between network layers, with an external pulse laser, which cannot be integrated on the same platform and the lack of amplification elements does not ease scalability for these networks. In our approach, we exploited semiconductor optical amplifier (SOA) as single-device weighting element as well as to compensate for on-chip losses, improve the weight tuning dynamic range. Most importantly, III–V material platform enables monolithic integration of the all-optical neurons (AONs) as well as all-optical neural networks (AONNs), by co-integrating multiple SOAs operating both in linear and non-linear regimes, and without hybrid connection between the two functionalities, taking advantage of on-chip rich nonlinearities such as cross-gain modulation (XGM), cross-phase modulation, four-wave-mixing (FWM). Furthermore, the SOA-based neural network can also be exploited in the near future for fast reconfiguration of the weights for ultra-fast optical training.

In previous research, we have realized an SOA-based linear unit for pattern classification [15], and further developed a monolithically integrated AON [16] for photonic deep neural networks, utilizing the SOAs operated in both linear and nonlinear regimes while processing the optical signal in the analogue domain. The all-optical approach is considered to reduce the number of expensive high-speed components and remove the distortions coming from E/O conversions while increasing the speed of consecutive linear and nonlinear processing in neural layers. In our AONs, we used WDM as neuron inputs, and the gain/loss from linear SOAs for assigning the weight factor, and finally the nonlinear SOA for multi-wavelength-to-monochromatic channel conversion based on XGM. Exploiting XGM in wavelength conversion is beneficial because of the simple device structure and the high passband width. However, the conversion with XGM will suffer from extinction ratio degradation. To overcome this problem, a two-stage and three-stage conversion [17, 18], with multi-time conversion from the split input signal to the continuous wave (CW) signal, has been proposed for zero power penalty conversion with stage related extinction ratio enhancement. In our case, we utilize XGM for multi-signal to single signal conversion, which corresponds to a kind of multi-stage conversion, but employing parallel optical inputs, instead of serial single signal multi-time conversions, aiming to achieve peak to peak signal enhancement too. With peak-to-peak signal enhancement (later defined as extinction ratio for analogue signal), it is possible to achieve a non-degraded signal processing of noisy inputs, enabling the connections of an arbitrary number of layers of neurons.

The noise model for in-line amplification transmission [19, 20] has been used in the communication system for many years, however, no model is available for optical neurons based on WDM channels and series of optical amplifiers working in the linear and non-linear regimes. To investigate the scalability of SOA-based AONNs, we need to emulate and validate the scaling of the AONN, starting from experimental results and developing a theoretical model to show the possibility of a larger-scale network.

In this paper, we demonstrate an emulation of the scaling of the integrated photonic deep neural network in terms of number of wavelengths and optical signal noise ratio (OSNR) evolution, where the possible number of wavelengths represents the scaling of the neuron connectivity and the OSNR evolution is used to interpret the scaling of the layer number. In section 2, we explain the experimental setup used to validate the model and extract the parameters to be used in the following sections. Thereafter in section 3, we present the theoretical method to simulate the response of the AONs with the experimental conditions. In section 4, both the experimental results and the simulations are shown, and the signal degradation is analysed to indicate the possibility of scaling condition of SOA-based AONN.

2. Experimental setup and emulation

For the emulation of the deep neural network scalability, we want to determine the performance of the neuron optical signal output with respect to the number of input channels as well as the network performance with the cascade of layers of neurons. 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 in the forward direction. 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 deep neural network. Moreover, for the layer scalability, we emulate the signal degradation defined by the OSNR at the input and output of each layer. However, the output OSNR at

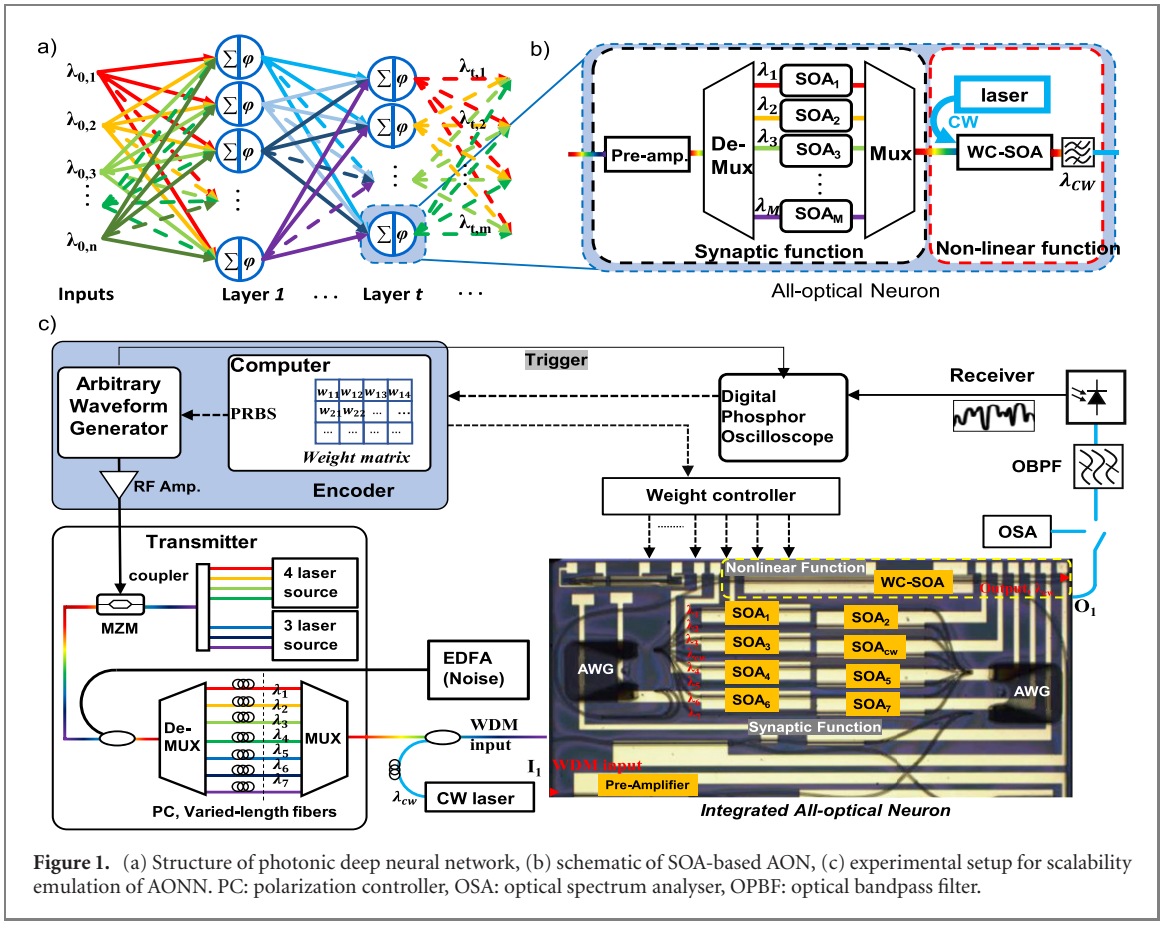

Figure 1. (a) Structure of photonic deep neural network, (b) schematic of SOA-based AON, (c) experimental setup for scalability emulation of AONN. PC: polarization controller, OSA: optical spectrum analyser, OPBF: optical bandpass filter.

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 emulate the scaling of the neural layers, which defines the depth of the all-optical deep 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 AON. Figure 1(a) presents the structure of the all-optical deep neural network with WDM channel interconnections consisted of multiple AONs per layer and feed forwarded to deeper layers. This fully-connected feed-forward multi-layer deep neural network is the network topology used in this paper. And we can obtain the weighting matrix from off-chip training and run the inference with the AONN as in [15, 16]. The circles represent the AONs with optical linear and nonlinear units for weighed addition $\sum$ and activation function φ, respectively. The Mth neuron in the tth layer (grey box) gives an output $y_{t,M} = \varphi\left(\sum w_i x_i\right)$, imprinted on $\lambda_{t,M}$, with w_i and x_i being the ith weighting factor and input data, respectively. The AONs receive parallel data in WDM channels from previous layer and generate single wavelength output, which is split and fully-connected to neurons in the next layer. The output of AONs may be multi-wavelength [21], which may reduce the loss by the splitter to the next layer, and the effect of the signal quality demands further investigation in the future. Since the network is fully programmable, it is always possible to use fewer neuron connections than in a fully-connected network, i.e., for pruned deep neural network, only selected inputs are fed to the next layer. In other words, the structure topology will still remain one of the feed-forward neural networks, but the connections may be enabled or not, depending on or off, at a higher control level. However, the nodes are not designed to be freely connected to any other node in the same layer. For a target neural network size bigger than the physical neural network, it is possible to implement the deep neural network by exploiting time-multiplexing of the input data and reconfiguring the weighting SOAs in the nanosecond regime. This demands further development of the control panel in the future.

Figure 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.

Figure 1(c) illustrates the experimental setup for assessing the performance of the SOA-based AON and emulation of the scalability of the AONN. The micrography with SOA schematic shows the fabricated all-optical integrated neuron used in this work. The footprint of the weighting SOA is 500 $\times$ 2 μm^2, which grows up to 500 $\times$ 100 μm^2 when we include the metal tracks and pads. The eight-channel neuron occupies 4.5 $\times$ 1.5 mm^2 area, including the wavelength converter based optical non-linear function. With the large-scale integration process as shown in [15], it is already possible to co-integrate 16 AONs with the proposed design. And the number of electric controls per complete neuron is defined by the number of the weighting SOAs and the wavelength conversion SOAs. For a layer of eight-input 16 neurons, 144 connections are required. The total integrated neuron number can grow up to more than 800, on a 3 inch InP wafer. In addition, a much more compact integration of active and passive elements can be achieved on the novel InP membrane on silicon platform [22]. Furthermore, more compact control connections may be obtained when routing the metal tracks at the die edge for ease of wire bonding or when using advanced flip-chip bonding.

As shown in the grey box in figure 1(b), the multiple inputs of the AON 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 seven WDM inputs, matching the on-chip passband wavelength defined by the arrayed waveguide grating 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^{-1} pseudo-random binary sequence on–off keying signal, generated by an arbitrary waveform generator (Tektronix, AWG7122B). Then the input is sent to the AON after decorrelation and multiplexing, with a power of −17.5 dBm per channel. Although we use on–off keying signal as input. A spiking input signal is also possible using the wavelength conversion scheme [23], this simple AON structure may be interesting for processing spiking optical input in the future. In this work, an external CW laser at 1546.72 nm is multiplexed to the WDM inputs to replace the tunable laser on-chip [24], 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 [25]. 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. We assess the converted signal 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 NRMSE between the measured output time traces and the expected time traces calculated based on 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 response 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.

3. Noise modelling and error evolution

To understand the performance of the AONN at large scale, we have developed a noise model 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 A):

$$\mathrm{NRMSE} = \sqrt{N_e}/S, \tag{1}$$

where N_e is the 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, we firstly derive an empirical relation between the spontaneous emission source noise and the error on the time traces (see appendix A):

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

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 equations (A14)–(A16) in appendix A. One can now determine the noise level for a given NRMSE value using the inverse relation of equation (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, we need to consider the noise accumulation along with its propagation in the AON. As shown in figure 1(a), the AON includes three 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 (B4) in appendix 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 FWM effect [16]. Since we estimate a carrier lifetime τ for the WC-SOA to be about 200 ps [26], the FWM effect is neglectable when detuning between the probe and pump channel is $\Delta f \gg 1/(2\pi\tau) \approx 1$ GHz. 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, we consider 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. With WDM inputs, the spontaneous emission density at the neuron output is defined as [27]:

$$\rho_{ASE} = \sum \eta_i \rho_{sse,i} + \sum \tilde{\eta}_i \rho_{wc-ASE,i}, \tag{3}$$

with

$$\eta_i = |P_i/(P_T w_i) \cdot F(L)|, \tag{4}$$

$$\tilde{\eta}_i = \eta_i / G, \tag{5}$$

$$w_i = p_i / p_T, \tag{6}$$

$$p_T = \sum p_i, \quad P_T = \sum P_i, \tag{7}$$

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 ith 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 ith input channel, respectively. Note in equation (3), the CW laser channel (0th 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 [28, 29] (see equation (B4) in appendix 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 (3) represents the noise conversion from the input noise and the second term shows the internal ASE contribution of the WC-SOA.

In equation (3), we need to know η_i, $\tilde{\eta}_i$, $\rho_{sse,i}$ and $\rho_{ASE,i}$. The conversion efficiency η_i and $\tilde{\eta}_i$ can be determine by working out $F(L)$ in equation (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 (B4).

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 [30]:

$$\rho_{\mathrm{ASE,c}} = h\nu_c\left\{n_{\mathrm{sp}}[G-1] + b_{\mathrm{sp}}\left(P_{\mathrm{T}}/P_{\mathrm{sat}}\right)\ln(G_0)\right\}, \tag{8}$$

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 ith channel λ_i can be approximated by a second-order polynomial in logarithm [31]:

$$\log(\rho_{\mathrm{ASE},i}) = \left(a_1(\lambda_i-\lambda_c)^2 + a_2(\lambda_i-\lambda_c) + 1\right)\log(\rho_{\mathrm{ASE,c}}), \tag{9}$$

where we have the maximum ASE at the centre wavelength. Equations (8) and (9) are the general noise models for SOAs, for respecting the ASE generation from pre-amplifier SOA, weight SOA, and WC-SOA. Hence, we can denote the ASE values for centre wavelength as $\rho_{\mathrm{pre-ASE,c}}$, $\rho_{\mathrm{w-ASE,c}}$ and $\rho_{\mathrm{wc-ASE,c}}$. And their corresponding ith ASE density as $\rho_{\mathrm{pre-ASE},i}$, $\rho_{\mathrm{w-ASE},i}$ and $\rho_{\mathrm{wc-ASE},i}$. Equation (8) can be obtained by measuring the ASE from WC-SOA as a function of the neuron input power curve, while equation (9) can be obtained by measuring the output optical spectrum. Therefore, the SSE noise at the WC-SOA input is:

$$\begin{aligned}\rho_{\mathrm{sse},i} = {} & \rho_{\mathrm{sse0},i}L_cG_1L_1G_iL_2 + G_iL_2\rho_{\mathrm{pre-ASE},i} \\ & + L_2\rho_{\mathrm{w-ASE},i},\end{aligned} \tag{10}$$

where $\rho_{\mathrm{sse0},i}$ are the original SSE noise at ith 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 ith weight SOA for individual weighting. Similarly, we get input signal power to the WC-SOA:

$$P_i = P_0L_cG_1L_1G_iL_2, \tag{11}$$

where P_0 is the input signal power per channel. Up to now, we can use equations (2)–(7) to obtain the ρ_{ASE} at the output. After filtering, the output noise power of the WC-SOA is:

$$P_{\mathrm{wc-SOA}} = L_{\mathrm{OBPF}}\rho_{\mathrm{ASE}}B_o, \tag{12}$$

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

The NRMSE can be estimated with the $N_{\mathrm{wc-SOA}}$ over the optical signal power, considering the conversion efficiency of the optical signal (see appendix C):

$$\mathrm{NRMSE} = \sqrt{(c_1'I_{\mathrm{wc-SOA}}^2 + c_2'I_{\mathrm{wc-SOA}} + c_3')}/\eta, \tag{13}$$

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 C, from equations (C7)–(C9), and $\eta = \sum\eta_i$, given by equation (C3).

With equations (3)–(13), we can determine the error evolution by tuning the input OSNR. And with equation (2) we can determine the equivalent output OSNR. Note that the input OSNR and the output NRMSE are directly measurable data in the setup described in section 2. The parameters used in the simulations are determined from the measurement results in the SOAs and photonic integrated AON 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, we determine the condition of the noise compression. 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, $P_{\mathrm{in}}/P_{\mathrm{sse}}$, we have an output OSNR, $P_{\mathrm{out}}/P_{\mathrm{ASE}}$, the input OSNR is

$$\mathrm{OSNR_{in}} = \frac{\sum P_i}{\sum P_{\mathrm{sse},i}} = \frac{\sum P_i}{B_o\sum\rho_{\mathrm{sse},i}}. \tag{14}$$

The output OSNR is

$$\mathrm{OSNR_{out}} = \frac{P_{\mathrm{out}}}{P_{\mathrm{ASE}}} = \frac{GP_{\mathrm{cw}}}{\rho_{\mathrm{ASE}}B_o}. \tag{15}$$

For a noise suppression:

$$\frac{\mathrm{OSNR_{in}}}{\mathrm{OSNR_{out}}} \leqslant 1\cdot \tag{16}$$

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

$$\mathrm{OSNR_{in}} \leqslant \frac{GP_{\mathrm{cw}} - \eta\overline{P}_{\mathrm{in}}}{\eta\overline{\rho}_{\mathrm{ASE}}/G + G\rho_{\mathrm{sse,\,c}} + \rho_{\mathrm{ASE,c}}} \cdot \frac{1}{B_{\mathrm{o}}}, \tag{17}$$

where P_{cw} is the power of CW laser, and $\overline{P}_{\mathrm{in}}$ is the average power of the input signal, G is the saturated gain of the WC-SOA. The equation (17) shows that it is possible to find an input OSNR for noise suppression if the $GP_{\mathrm{cw}} > \eta\overline{P}_{\mathrm{in}}$, and $\overline{\rho}_{\mathrm{ASE}}$, $\rho_{\mathrm{ASE,c}}$ is defined by the noise generation from the WC-SOA as shown in equations (8) and (9), and the noise from CW laser $\rho_{\mathrm{sse,c}}$ is either defined by equation (10) in this experiment, or by the SSE of a direct coupled CW-laser to the WC-SOA [16]. If the input OSNR satisfies the requirement in equation (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) when input signals are equalized in power. The condition of equation (17) for un-equalized WC-SOA inputs can be obtained numerically with equations (3), (13) and (16).

4. Experimental and simulation results

4.1. SOA characterisation

To determine the total ASE $\rho_{\mathrm{ASE,c}}$ after WC-SOA in the converted output signal using equations (1) and (2), we need to obtain the conversion efficiency η_i, the input SSE $\rho_{\mathrm{sse},i}$ and output ASE $\rho_{\mathrm{ASE},i}$. 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. To correctly represent the properties of the SOAs on chip, we use the on-chip SOAs as photodetectors 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. Figure 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 $P_{\mathrm{sat}} = 6$ dBm, and internal loss $\alpha' = 0.5$, using equation (B4) in appendix 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 $n_{\mathrm{sp}} = 7.3$ and $b_{\mathrm{sp}} = 1$ in equation (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 $G_{0,\text{pre-SOA}} = 9.5$ dB, saturation power $P_{\mathrm{sat,pre\text{-}SOA}} = 7.8$ dBm, normalized waveguide loss $\alpha'_{\text{pre-SOA}} = 0.6$, $n_{\mathrm{sp,pre\text{-}SOA}} = 3.6$ and $b_{\mathrm{sp}} = 1$. The obtained parameters are now used for the NRMSE estimation. The rest of the other parameters used in the simulations is listed in table 1.

4.2. Error evolution of SOA-based all-optical neuron

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 figure 3(a), the blue lines depict the recorded time traces at the output of the seven-channel AON after filtering. These are superimposed to the expected output time traces shown in red lines. The calculated NRMSE from seven-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 mA, 120 mA, 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 figure 3(b), 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 $\geqslant$−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 (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

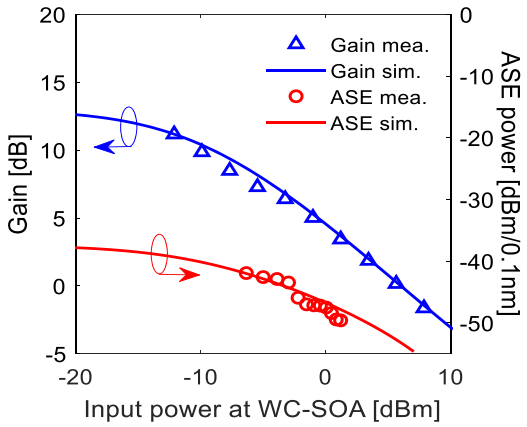

Figure 2. The gain (blue) and ASE power (red) vs input power at WC-SOA, from experiments (scatters) and simulation (solid line).

relation between the NRMSE 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 subsection 4.1), we can now simulate the error evolution of the AON. The blue, red, green, magenta scatters and solid lines in figure 3(b) illustrate the measured and simulated errors when tuning the OSNR at the neuron input from 15 dB to 45 dB, for an AON with two-, three-, four- and seven-channel inputs. The five- and six-channel curves are not shown as they are very close to the error trend for seven-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 figure 3(b) 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 figure 3(a). There is a slight discrepancy between the modelling curve and the experimental data, attributing to the assumption that the WDM is dense and the gain for different channel is identical so the averaged gain for different channel number input is used in the simulation. However, in the experiments, the gain for different channel is not identical. Specifically for a small number of channels, for instance, the case of two-channel weighted addition, the average gain, therefore, the noise, seen from equation (8), of the input signal is slightly higher in the experiment than the one used in the simulations, which causes the experimental data appears to be above the simulation data. The same model could be further improved by considering the gain spectrum and by handling it with large signal analysis [31]. None the less, 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 (17), there will be a OSNR that leads to $\mathrm{OSNR_{out}} \geqslant \mathrm{OSNR_{in}}$, 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 equations (3) and (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 figure 3(c) illustrates the output OSNR when tuning the input OSNR from 15 to 45 dB for two, three, four, and seven-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 (2) with given NRMSE from the data for corresponding input OSNR, as shown in figure 3(b). Using the same method, we can determine the output OSNR for all the input channels. The noise suppression effect

Table 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
$G_{0,\text{ pre-SOA}}$	Unsaturated gain, pre/w-SOA	9.5	dB
P_{sat}	Saturation power, WC-SOA	6	dBm
$P_{sat,pre/w\text{-}SOA}$	Saturation power, pre/w-SOA	7.8	dBm
α'	Normalized waveguide loss, WC-SOA	0.5	
$\alpha'_{\text{ pre/w-SOA}}$	Normalized waveguide loss, pre/w-SOA	0.6	
n_{sp}	Noise inversion parameter, WC-SOA	7.3	
$n_{sp,pre/w\text{-}SOA}$	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
$a_{1,pre/w\text{-}SOA}$	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}
$a_{2,pre/w\text{-}SOA}$	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/0.1 nm
P_0	Input signal power per channel in experiments	−17.5	dBm

is happened when input OSNR is smaller than 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 NRMSE at the AON output has been obtained as shown in figure 3(b) as a function of the input OSNR at the first layer. Given an input OSNR, we then get the output OSNR as shown in figure 3(c), 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.

4.3. Error evolution for the AON-based network

In figure 3(b) 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 figure 3(c). In figure 4(a), the solid lines and dashed lines plot the error and the OSNR evolutions, respectively, when cascading neural layers with four, five, six and seven neurons per layer (which is the same as saying with AON with four-, five-, six- and seven-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 minimum noise levels are defined by the crossing point shown in figure 3(b): 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 (2)). The error decrease then increased when increasing the input channel number from four- to seven-channels, which is attributing to the compromise of conversion efficiency and the gain at

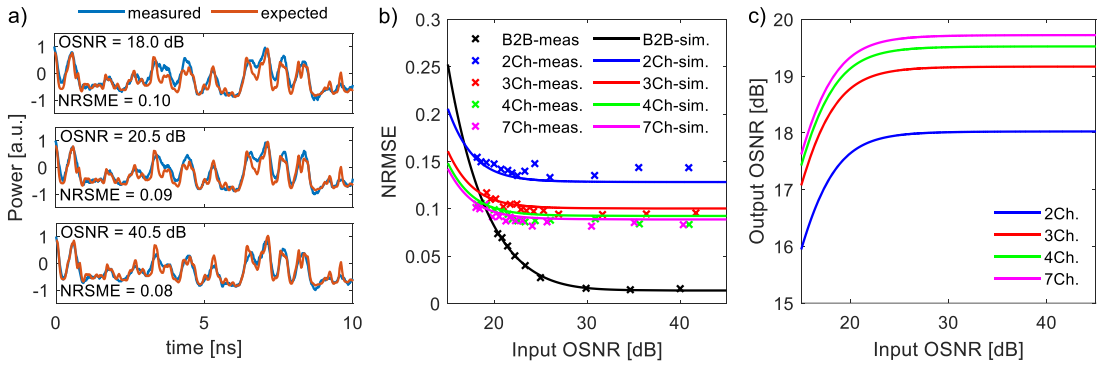

Figure 3. (a) The times measured (blue lines) at the output of AON with seven channel input, and the input OSNR setting at 18.0, 20.5, and 40.5 dB, compared to the expected (red lines) time traces calculated from weighed addition. (b) The measured (crosses) and simulated (solid line) NRMSE from the detection, for the B2B measurement at the photodetector (black), and AON with two, three, four, and seven-channel inputs. (c) The equivalent output OSNR (blue) and the noise figure (red) when tuning the input OSNR for two, three, four, and seven-channel AON.

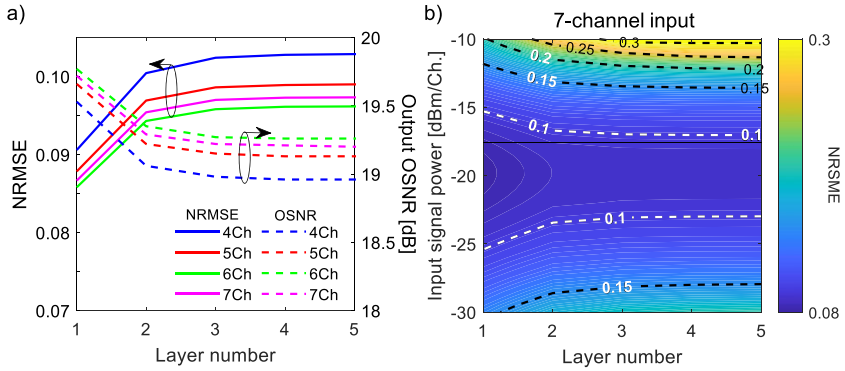

Figure 4. (a) The NRMSE (solid lines) and output OSNR (dashed lines) at the output of AON with four, five, six, and seven channel inputs as a function of the layer number. (b) The NRMSE changes when tuning the input signal power from −30 to −10 dBm per channel, and changing the layer number from one to five layers, for an AONNs with seven-channel inputs. The black line presents the magenta line in (a).

WC-SOA, known from equation (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 figure 2.

We further investigate the input power to the WC-SOA for the seven-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 figure 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 seven-channels input AON in figure 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 seven 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). Figure 5(a) depicts the NRMSE 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 $\leqslant$ 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.

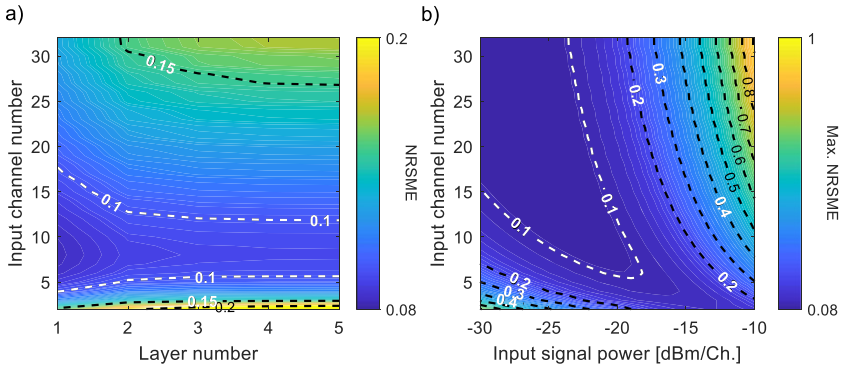

Figure 5. (a) The NRMSE changes vs input channel numbers and layer numbers for fixed input signal power at −20 dBm per channel. (b) The maximum NRMSE vs the input channel numbers when tuning the input signal power from −30 to −10 dBm per channel.

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. Figure 5(b) illustrates the maximum NRMSE of the AON for arbitrary layer number connection, obtained from the crossing point from figure 3(b), 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.

5. Conclusions

We emulate the scaling of AONN with photonic integrated SOA-based AON, utilizing XGM as nonlinear transfer function, and we develop the noise model 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 M:1 conversion on a single output, with 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 AONN, 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 M: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.

Acknowledgments

This work is financially supported by the Netherlands Organization of Scientific Research (NWO) under the Zwaartekracht programma, 'Research Centre for Integrated Nanophotonics', Grant No. 024.002.033.

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

Appendix A. Back-to-back measurement

Here we derive the format of the NRMSE vs input noise with fixed input power. The definition of the NRMSE in this paper is:

$$\mathrm{NRMSE} = \frac{\sqrt{\sum_{i=1}^{m}(x_i - E_i)^2/m}}{I_{\mathrm{s,max}} - I_{\mathrm{s,min}}}, \tag{A1}$$

where x_i is the ith measured data value and E_i is the ith expected value for input reference with calibrated weighted addition. m is the number of recorded data points. $I_{\mathrm{s,max}}$ and $I_{\mathrm{s,min}}$ are the maximum and minimum photocurrent on the detection. Considering the noise at the photodetector has a Gaussian distribution $\mathcal{N}(0, N_{\mathrm{e}})$, with N_{e} as the electrical noise at the receiver, equation (A1) becomes:

$$\mathrm{NRMSE} = \sqrt{N_{\mathrm{e}}}/(I_{\mathrm{s,max}} - I_{\mathrm{s,min}}). \tag{A2}$$

we get equation (1) with substituting $S = (I_{\mathrm{s,max}} - I_{\mathrm{s,min}})$ 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 [32]:

$$N_{\mathrm{shot}} = 2eB_{\mathrm{e}}\left(I_{\mathrm{s}} + I_{\mathrm{sp}}\right), \tag{A3}$$

$$N_{\mathrm{s-sp}} = 4I_{\mathrm{s}}I_{\mathrm{sp}}B_{\mathrm{e}}/B_{\mathrm{o}}, \tag{A4}$$

$$N_{\mathrm{sp-sp}} = I_{\mathrm{sp}}^2 B_{\mathrm{e}}(2B_{\mathrm{o}} - B_{\mathrm{e}})/B_{\mathrm{o}}^2, \tag{A5}$$

$$N_{\mathrm{th}} = I_{\mathrm{th}}^2, \tag{A6}$$

$$N = N_{\mathrm{shot}} + N_{\mathrm{s-sp}} + N_{\mathrm{sp-sp}} + N_{\mathrm{th}}, \tag{A7}$$

where the noise N consists of shot noise N_{shot}, signal–spontaneous beating noise $N_{\mathrm{s-sp}}$, spontaneous–spontaneous beating noise $N_{\mathrm{sp-sp}}$, 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:

$$I_{\mathrm{s,max\,/\,min}} = R_0 P_{\mathrm{s,max\,/\,min}}, \qquad I_{\mathrm{sp}} = R_0 P_{\mathrm{sp}} F_{\mathrm{e}}, \tag{A8}$$

with R_0 is the responsivity of the photodiode. In case of APD and followed with an electrical amplifier, $R_0 = M_{\mathrm{apd}}R_{\mathrm{apd}}G_{\mathrm{Rx}}$, where M_{apd}, R_{apd} are the multiplication factor and responsivity of the APD when $M_{\mathrm{apd}} = 1$ 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:

$$I_{\mathrm{s,max}} = rI_{\mathrm{s,min}}, \tag{A9}$$

$$I_{\mathrm{s}} = \left(I_{\mathrm{s,max}} + I_{\mathrm{s,min}}\right)/2, \tag{A10}$$

yields,

$$P_{\mathrm{s}} = (P_{\mathrm{s,max}} + P_{\mathrm{s,min}})/2, \tag{A11}$$

$$S = 2I_{\mathrm{s}}(r-1)/(r+1). \tag{A12}$$

From equation (A2),

$$\mathrm{NRSME} = \sqrt{N_{\mathrm{e}}}/S\cdot \tag{A13}$$

Substituting (A8) and (A11), with fixed r and I_{s}, we get equation (2):

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

with,

$$c_1 = B_e(2B_o - B_e)/(B_o^2 S^2), \tag{A14}$$

$$c_2 = 2\left(eB_e + 2I_s B_e/B_o\right)/S^2, \tag{A15}$$

$$c_3 = \left(2eB_e I_s + I_{th}^2\right)/S^2. \tag{A16}$$

Appendix B. Conversion efficiency

Here we derive the calculation of conversion efficiency at the WC-SOA, 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 ith input signal at WC-SOA. For cross gain modulation, the conversion from jth input channel to ith 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 [29]:

$$\begin{aligned}\eta_{ki} &= \left|\frac{p_k(L)}{P_k(L)} \Big/ \frac{p_i(0)}{P_i(0)}\right| \\ &= \left|\frac{p_k(0)P_i(0)}{P_k(0)p_i(0)} - \frac{p_T(0)P_i(0)}{P_T(0)p_i(0)}F(L)\right|,\end{aligned} \tag{B1}$$

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

$$F(L) = 1 - e^{-K(L)}, \tag{B2}$$

$$K(L) = \frac{1}{1 - j\omega\tau\alpha'}\left\{\alpha' \ln\frac{G_0}{G} - \ln\left[1 - \frac{(G-1)P_T(0)/P_{sat}}{1 + GP_T(0)/P_{sat} + j\omega\tau}\right]\right\}, \tag{B3}$$

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:

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

The amplifier saturated gain is defined from [19, 29]:

$$\alpha' \ln\frac{G_0}{G} = \ln\left\{\frac{1 - \alpha'\left[1 + P_T(0)/P_{sat}\right]}{1 - \alpha'\left[1 + GP_T(0)/P_{sat}\right]}\right\}. \tag{B4}$$

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

$$r' = \frac{p_i(0)}{P_i(0)}, \quad \text{for } i = 1, \ldots, M. \tag{B5}$$

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

$$r' = 2(r-1)/(r+1). \tag{B6}$$

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

$$w_i' = \frac{p_i(0)}{p_T(0)}. \tag{B7}$$

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

$$\begin{aligned}\eta_{ci} &= \left|\frac{p_c(L)}{P_c(L)} \Big/ \frac{p_i(0)}{P_i(0)}\right| \\ &= \left|\frac{p_T(0)P_i(0)}{P_T(0)p_i(0)}F(L)\right| \\ &= \left|\frac{P_i(0)}{P_T(0)w'_i}F(L)\right|.\end{aligned} \tag{B8}$$

Denoting $\eta_i = \eta_{ci}$ in equation (B4) and define $P_i(0)$ at the input as P_i, we obtain equation (2) in the main text.

Appendix C. NRMSE versus input noise

Here we derive the format of the output NRMSE vs input noise with amplified converted signal and changed extinction ratio.

The modulation index of the output after conversion, for the amount from the ith input channel, from equation (B1), is:

$$\frac{p_{c,i}(L)}{P_c(L)} = \eta_i \frac{p_i(0)}{P_i(0)}. \tag{C1}$$

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

$$\begin{aligned} r_c' &= \frac{p_c(L)}{P_c(L)} = \sum_{i=1}^{M} \frac{p_{c,i}(L)}{P_c(L)} \\ &= r' \sum_{i=1}^{M} \eta_i. \end{aligned} \tag{C2}$$

So that the total conversion efficiency is:

$$\eta = r_c'/r' = \sum_{i=1}^{M} \eta_i. \tag{C3}$$

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:

$$\begin{aligned} S_c &= 2R_0P_c(L)r_c' = S_0P_c(L)r_c'/(P_{s0}r') \\ &= \eta S_0P_c(L)/P_{s0} = \eta S_0G'. \end{aligned} \tag{C4}$$

If $G' = P_c(L)/P_{s0} = 1$, i.e., the same average power at the output, the NRMSE 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:

$$I_{sp,c} = G'I_{sp} \tag{C5}$$

$$\begin{aligned} I_{s,c} &= R_0P_c(L) = R_0G'P_{s0} \\ &= G'I_s \end{aligned} \tag{C6}$$

substitute (C4)–(C6) in equation (2) with (A14)–(A16) the NRMSE will be:

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

With

$$c_1' = B_e(2B_o - B_e)/(B_o^2S_0^2), \tag{C7}$$

$$c_2' = 2\left(B_e/G' + 2I_sB_e/B_o\right)/S_0^2, \tag{C8}$$

$$c_3 = \left(2B_eG'I_s + I_{th}^2\right)/\left(G'S_0\right)^2. \tag{C9}$$

Substituting $I_{sp,c}$ with $I_{wc\text{-}SOA}$, we can obtain equation (10).

Appendix 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 equations (13)–(16), we have,

$$\frac{GP_{cw}}{\rho_{ASE}B_o} \geqslant \frac{P_{in}}{B_o\rho_{in}} = \frac{\sum P_i}{B_o \sum \rho_{sse,i}}. \tag{D1}$$

From equation (3), unfold the CW channel, it yields,

$$\frac{GP_{\mathrm{cw}}}{\sum \eta_i \rho_{\mathrm{sse},i} + \sum \eta_i \rho_{\mathrm{wc-ASE},i}/G + G\rho_{\mathrm{sse,\,cw}} + \rho_{\mathrm{ASE,cw}}} \geqslant \frac{\overline{P}_{\mathrm{in}}}{\overline{\rho}_{\mathrm{sse}}}. \tag{D2}$$

With $\overline{P}_{\mathrm{in}} = \frac{1}{M}\sum P_i$, $\overline{\rho}_{\mathrm{sse}} = \frac{1}{M}\sum \rho_{\mathrm{sse},i}$, 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 $\overline{P}_{\mathrm{in}}$. With $\overline{\rho}_{\mathrm{ASE}} = \frac{1}{M}\sum \rho_{\mathrm{ASE},i}$, after some algebra, we obtain,

$$\overline{\rho}_{\mathrm{sse}} \geqslant \frac{\eta\overline{\rho}_{\mathrm{ASE}}/G + G\rho_{\mathrm{sse,cw}} + \rho_{\mathrm{ASE,cw}}}{\frac{GP_{\mathrm{cw}}}{\overline{P}_{\mathrm{in}}} - \eta}. \tag{D3}$$

Therefore,

$$\frac{\overline{P}_{\mathrm{in}}}{\overline{\rho}_{\mathrm{sse}}B_{\mathrm{o}}} \leqslant \frac{GP_{\mathrm{cw}} - \eta\overline{P}_{\mathrm{in}}}{\eta\overline{\rho}_{\mathrm{ASE}}/G + G\rho_{\mathrm{sse,cw}} + \rho_{\mathrm{ASE,cw}}} \cdot \frac{1}{B_{\mathrm{o}}}. \tag{D4}$$

With $\mathrm{OSNR}_{\mathrm{in}} = \frac{\overline{P}_{\mathrm{in}}}{\overline{\rho}_{\mathrm{sse}}B_{\mathrm{o}}}$, the equation (17) is obtained. And from equation (D3), we know equation (D4) exits only if $\frac{GP_{\mathrm{cw}}}{\overline{P}_{\mathrm{in}}} \geqslant \eta$.

ORCID iDs

Bin Shi https://orcid.org/0000-0003-3005-685X
Ripalta Stabile https://orcid.org/0000-0001-5197-3150

References

[1] Kendall J D and Kumar S 2020 The building blocks of a brain-inspired computer *Appl. Phys. Rev.* **7** 011305
[2] Choquette J, Gandhi W, Giroux O, Stam N and Krashinsky R 2021 NVIDIA A100 tensor core GPU: performance and innovation *IEEE Micro* **41** 29–35
[3] Jouppi N P *et al* 2017 In-datacenter performance analysis of a tensor processing unit *Proc.—Int. Symp. on Computer Architecture* vol part F1286 (New York: ACM) pp 1–12
[4] Akopyan F *et al* 2015 TrueNorth: design and tool flow of a 65 mW 1 million neuron programmable neurosynaptic chip *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **34** 1537–57
[5] Furber S B, Galluppi F, Temple S and Plana L A 2014 The SpiNNaker project *Proc. IEEE* **102** 652–65
[6] Pei J *et al* 2019 Towards artificial general intelligence with hybrid Tianjic chip architecture *Nature* **572** 106–11
[7] Kitayama K I, Notomi M, Naruse M, Inoue K, Kawakami S and Uchida A 2019 Novel frontier of photonics for data processing-photonic accelerator *APL Photon.* **4** 90901
[8] Shen Y *et al* 2017 Deep learning with coherent nanophotonic circuits *Nat. Photon.* **11** 441–6
[9] Hamerly R, Sludds A, Bernstein L, Soljačić M and Englund D 2018 Large-scale optical neural networks based on photoelectric multiplication *Phys. Rev.* X **9** 021032
[10] Mourgias-Alexandris G, Totovic A, Tsakyridis A, Passalis N, Vyrsokinos K, Tefas A and Pleros N 2020 Neuromorphic photonics with coherent linear neurons using dual-IQ modulation cells *J. Lightwave Technol.* **38** 811–9
[11] Tait A N, Wu A X, de Lima T F, Zhou E, Shastri B J, Nahmias M A and Prucnal P R 2016 Microring weight banks *IEEE J. Sel. Top. Quantum Electron.* **22** 312–25
[12] Shastri B J, Tait A N, de Lima T F, Nahmias M A, Peng H-T and Prucnal P R 2017 Principles of neuromorphic photonics *Advanced Photonics 2017 (IPR, NOMA, Sensors, Networks, SPPCom, PS)* p PTu3C.4
[13] Feldmann J *et al* 2021 Parallel convolutional processing using an integrated photonic tensor core *Nature* **589** 52–8
[14] Feldmann J, Youngblood N, Wright C D, Bhaskaran H and Pernice W H P 2019 All-optical spiking neurosynaptic networks with self-learning capabilities *Nature* **569** 208–14
[15] Shi B, Calabretta N and Stabile R 2020 Deep neural network through an InP SOA-based photonic integrated cross-connect *IEEE J. Sel. Top. Quantum Electron.* **26** 1–11
[16] Shi B, Calabretta N and Stabile R 2022 InP photonic integrated multi-layer neural networks: architecture and performance analysis *APL Photon.* **7** 10801
[17] Simon J C, Lablonde L, Valiente I, Billes L and Lamouler P 1995 Two-stage wavelength converter with improved extinction ratio *Optical Fiber Communications Conf.* (Optica Publishing Group) p PD15
[18] Mao Y, Lu Z G, Chrostowski J, Hong J and Misner R 1999 Three-stage wavelength converter based on cross-gain modulation in semiconductor optical amplifiers *Opt. Commun.* **167** 57–66
[19] Connelly M J 2002 *Semiconductor Optical Amplifiers* (Dordrecht: Kluwer)
[20] Baney D M, Gallion P and Tucker R S 2000 Theory and measurement techniques for the noise figure of optical amplifiers *Opt. Fiber Technol.* **6** 122–54
[21] Raz O, Herrera J, Calabretta N, Tangdiongga E, Anantathanasarn S, Nötzel R and Dorren H J S 2008 Non-inverted multiple wavelength converter at 40 Gbit/s using 1550 nm quantum dot SOA *Electron. Lett.* **44** 988
[22] Jiao Y *et al* 2020 Indium phosphide membrane nanophotonic integrated circuits on Silicon *Phys. Status Solidi a* **217** 1–12
[23] Mourgias-Alexandris G, Tsakyridis A, Passalis N, Tefas A, Vyrsokinos K and Pleros N 2019 An all-optical neuron with sigmoid activation function *Opt. Express* **27** 9620–30

[24] Shi B, Prifti K, Magalhães E, Calabretta N and Stabile R 2020 Lossless monolithically integrated photonic InP neuron for all-optical computation *Optical Fiber Communication Conf. (OFC) 2020 (2020)* (The Optical Society) p W2A.12
[25] Yao W, Gilardi G, D'Agostino D, Smit M K and Wale M J 2017 Monolithic tunable coupled-cavity WDM transmitter in a generic foundry platform *IEEE Photon. Technol. Lett.* **29** 496–9
[26] Usami M, Tsurusawa M and Matsushima Y 1998 Mechanism for reducing recovery time of optical nonlinearity in semiconductor laser amplifier *Appl. Phys. Lett.* **72** 2657
[27] Obermann K, Koltchanov I, Petermann K, Diez S, Ludwig R and Weber H G 1997 Noise analysis of frequency converters utilizing semiconductor-laser amplifiers *IEEE J. Quantum Electron.* **33** 81–8
[28] Davies D A O 1995 Small-signal analysis of wavelength conversion in semiconductor laser amplifiers via gain saturation *IEEE Photon. Technol. Lett.* **7** 617–9
[29] Mecozzi A 1996 Small-signal theory of wavelength converters based on cross-gain modulation in semiconductor optical amplifiers *IEEE Photon. Technol. Lett.* **8** 1471–3
[30] Obermann K, Kindt S, Breuer D, Petermann K, Schmidt C, Diez S and Weber H G 1997 Noise characteristics of semiconductor-optical amplifiers used for wavelength conversion via cross-gain and cross-phase modulation *IEEE Photon. Technol. Lett.* **9** 312–4
[31] Obermann K, Kindt S, Breuer D and Petermann K 1998 Performance analysis of wavelength converters based on cross-gain modulation in semiconductor-optical amplifiers *J. Lightwave Technol.* **16** 78–85
[32] Olsson N A 1989 Lightwave systems with optical amplifiers *J. Lightwave Technol.* **7** 1071–82