Error Analysis of a 3-Layer SOA-based Photonic Deep Neural Network for Image Classification

B. Shi, N. Calabretta, and R. Stabile

Institute for Photonic Integration, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands

We demonstrate Iris flower classification with a trained 3-layer photonic deep neural network, which is implemented by a monolithically integrated SOA-based InP cross-connect chip, achieving a final accuracy of 85.8%. An in-depth analysis of the error evolution in the system suggests that the E/O conversions from layer to layer contribute to the majority of the error.

Introduction

Deep learning neural networks are attracting a huge interest owing to their versatile applications in data features extraction, image classification, time series prediction, and system optimization. Micro-electronics has been developed to decrease power consumption with neuromorphic approaches, which is inspired by the signal processing structure in the brain. However, the processing speed of the information is constrained to few GHz due to the limited bandwidth of the electrical interconnections between neurons. Photonic neural networks are being proposed to outperform the interconnectivity bandwidth in electronics. Optical neural networks with bulky discrete components have been demonstrated [1], but path phase differences increase complexity. Photonic deep neural network (PDNN) has been proposed based on a coherent approach using Mach-Zehnder Interferometer elements [2], but this suffers phase noise and noise accumulation as it consists of several stages for realizing one single layer of neurons. Banks of micro-ring resonators for weighting have been proposed with wavelength division multiplexing (WDM) inputs to take advantage of the optical interconnection bandwidth [3], but thermal crosstalk and low dynamic range complicate the weight calibration. We have recently proposed an InP SOA-based photonic integrated circuit (PIC) for enabling WDM connectivity and a wider dynamic range for PDNN [4].

In this work, we exploit an InP cross-connect chip, based on semiconductor optical amplifier (SOA) and array waveguide gratings (AWGs), as one layer of a PDNN. By setting the gain of the SOA as trained weighted factor for the WDM input, the cross-connect chip is employed as one layer of neurons in PDNN with off-line nonlinear function. By feeding the layer output back to the optical input, and by reconfiguring the on-chip weight matrix, a 3-layer PDNN is performed, to demonstrate the Iris flower classification [5]. Here, we focus on the analysis of the error evolution along with the signal processing from one layer to the next layer to investigate the major contributions.

Experiments and results

Fig.1a shows the experimental setup for image classification. The photonic integrated cross-connect is used as one layer of 4 neurons. Fig.1b sketches the weighted addition performed within a neuron and the corresponding photonic scheme. The input data are encoded within a WDM source with different channels, which will be de-multiplexed and weighted with the desired gain as a weight factor, then combined at the output as a summation of all the signals. The yellow boxes on the PIC in Fig. 1a show the used 4 neurons on-chip. The nonlinear function is not co-integrated but realized off-line via a computer. For the Iris flower classification problem, the database is made of three classes (Setosa, Versicolor, and Virginica) of 50 instances each. Per each instance, the iris flower category is identified by observing four of its attributes (length and width of sepals and petals). The training of the 3-layer deep neural network (DNN), shown in Fig. 1c, with hyperbolic tangent function as nonlinear function, is done on the simulation platform Tensorflow [6]. 120 instances are used as training database in this demonstration.

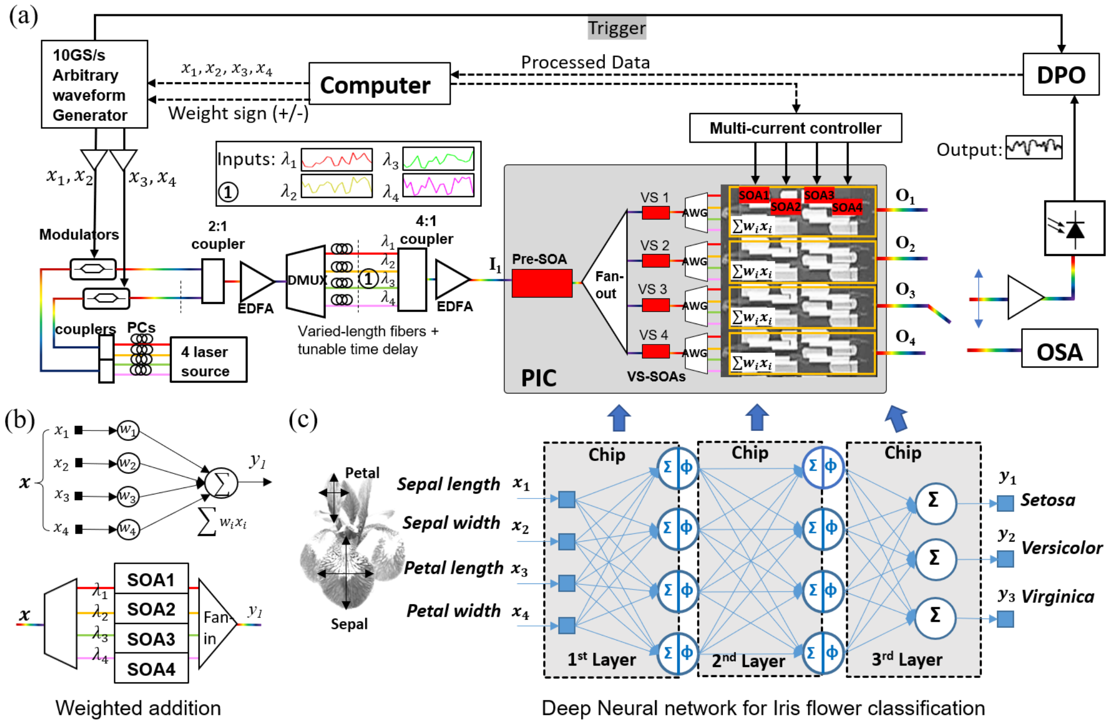

Fig. 1 (a) The experimental setup. (b) Weighted addition in the artificial neuron model and the corresponding photonic implementation. (c) Three-layer deep neural network for image classification, with reconfigurable photonic layer in gray box.

The Iris features are encoded with 2^6 levels, through an arbitrary waveform generator at 10 GSample/s and modulated on 4 channel WDM source with wavelengths at λ_1=1539.83 nm, λ_2=1543.10 nm, λ_3=1546.26 nm, and λ_4=1549.06 nm. The inputs are synchronized by a time delay stage, amplified, and coupled to the PIC via a lensed fiber. The weight SOAs are controlled by a multi-current controller. And the positive/negative weighting implementation is reported in [5]. The summation of the weighted output signal is accessed with a scanned lensed fiber and detected by a pre-amplified AC-coupled APD and a digital phosphor oscilloscope. The data is then processed with the computer as the output of a neuron layer. We calculate the final accuracy for the 4 cases: (a) the trained DNN on TensorFlow; (b) use of the chip as the 1st layer of the DNN only; (c) use the chip as 1st and 2nd layer of the DNN, after reconfiguring the on-chip weights; (d) use the chip three times as 1st, 2nd and 3rd layer of the DNN, by reconfiguring the on-chip weights. Table 1 lists the prediction accuracy and the corresponding correlation matrix between the predictions and labels when implementing the photonic neuron layers in the DNN. The accuracy decreases along the information processing when adding photonic layers, from the simulated DNN accuracy of 95% to the final accuracy of 85.8%.

Table 1 Prediction accuracy as a function of the involved photonic layers (PL)

Implementation	Trained DNN			PL 1			PL1 & 2			PL1, 2, & 3		
Accuracy	95%			91.7%			87.5%			85.8%		
Labels / Prediction	Seto.	Vers.	Virg.	Seto.	Vers.	Virg.	Seto.	Vers.	Virg.	Seto.	Vers.	Virg.
Setosa	42	0	0	42	1	0	42	1	0	42	1	0
Versicolor	0	30	0	0	26	0	0	23	2	0	24	5
Virginica	0	6	42	0	9	42	0	12	40	0	11	37

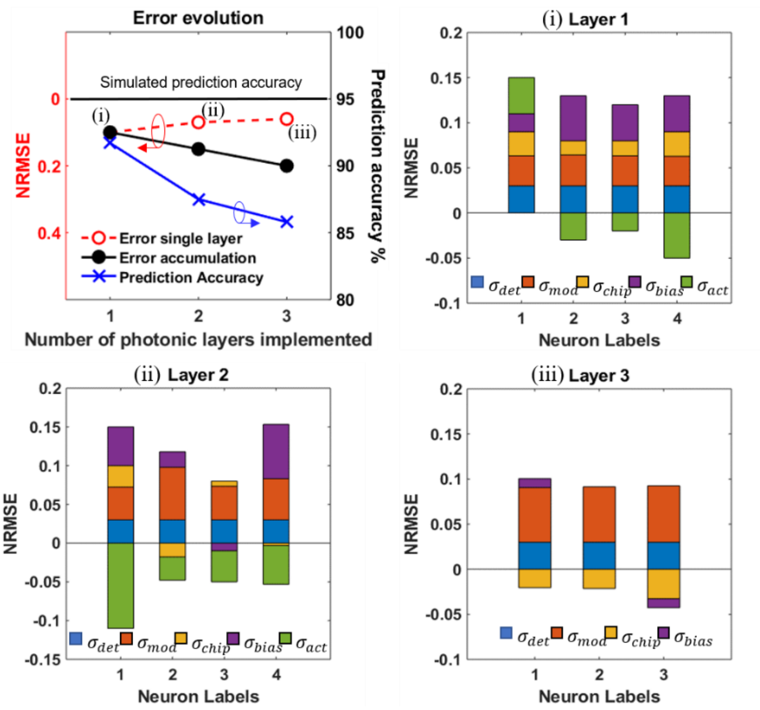

Fig. 2 Top left: the error evolution and prediction accuracy vs. number of photonic layers implemented. The error contribution analysis in (i) Layer 1, (ii) Layer 2, and (iii) Layer 3, resulting in an average error of single layer shown on top left.

The normalized root mean square error (NRMSE) is calculated to represent the distortion induced by one stage. Figure 2 illustrates error evolution when the signal is processed along the 3-layer PDNN. The filled black circles represent the accumulated error at the output port comparing with the expectation. The open red circles show the error induced by the single layers, and the blue crosses plot the prediction accuracy as calculated in Table 1. We could see the accumulated error increasing from 0.1 to 0.2 as new photonic layers are added. The error induced in a single layer is flat with values around 0.08, which suggests that the interface from layer to layer plays an important role in error accumulation.

To see the detailed contribution of different components in one layer of neurons, we consider the impairments from signal detection, data modulation, chip operation, bias and nonlinear function for each layer. For the detection error, that is obtained by calculating the maximum error between the detected data and the average data from N times detection. The measured errors range from 0.015 to 0.026 over 20 times of measurements. This error is set to a maximum fixed value of 0.03 and assumed to be identical for all the wavelengths. The error from the modulation is measured by comparing the original data and the detected data at the input when a 0.5 nm bandpass filter is applied on the individual channel. The contribution to the error of the modulation at the layer output is calculated by the weighted summation of the modulation errors from different channels, i.e. modulation error at output of the *i*th layer is $\boldsymbol{\sigma}_{Li} = \sum \boldsymbol{W}_{Li,} \cdot \boldsymbol{\sigma}_{L(i-1)}$, where $\boldsymbol{W}_{Li}$ is the weight matrix in layer *i*, and $\boldsymbol{\sigma}_{L(i-1)}$ is the vector of errors from the previous layer. The error from the output of the chip is accessed by comparing the output data to the expectation with the input data after subtracting the error from detection and modulation. Furthermore, error from bias and hyperbolic tangent function is obtained by calculating the additional error added to the previous stage.

Fig 2 (i) to (iii) show the error contributions from the different stages in layer 1, 2, and 3, corresponding to the average error at each single layer labeled (i) Layer 1, (ii) Layer 2, and (iii) Layer 3 on the top left figure. From the results, we could see that the hyperbolic tangent activation function reduces the error pronouncedly, (see green bars) in Fig 2(i), Neuron 2 to 3 in Layer 1, and in Fig 2(ii) Neurons in layer 2, as the output data are lying on the saturation region where the activation function compresses the data within (-1, 1) regime. The activation function induces an extra error in the case of Neuron 1 of Layer 1 since the output data from the chip are close to zero, whose error will easily enhance after the activation function. The error induced by the PIC is not the main contributor, while the E/O conversion is the major contributing factor to the error. Reducing the error from the E/O conversion is expected to improve the performance of the photonic deep neural network.

Conclusion

We have successfully demonstrated a 3-layer photonic neural network with the Iris flower classification problem. The final prediction is only 9.5% less than the trained DNN on the computer with an NRMSE of 0.2 at the output layer. This suggests that the photonic DNN is robust to the impairment. The complete error contribution analysis shows that the E/O conversion is the main factor of signal distortion along with the signal processing: an all-optical deep neural network is expected to improve final performance.

Acknowledgment

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

References

[1] J. Bueno *et al.*, "Reinforcement learning in a large-scale photonic recurrent neural network," *Optica*, vol. 5, no. 6, p. 756, Jun. 2018.
[2] Y. Shen *et al.*, "Deep learning with coherent nanophotonic circuits," *Nat. Photonics*, vol. 11, no. 7, pp. 441–446, Jun. 2017.
[3] 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. 3427–3439, 2014.
[4] 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.
[5] 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, pp. 1–11, Jan. 2020.
[6] M. Abadi *et al.*, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," Mar. 2016.