European Conference on Optical Communication (ECOC) 2022 © Optica Publishing Group 2022

High-speed analog photonic computing with tiled matrix multiplication and dynamic precision capabilities for DNNs

George Giamougiannis[(1)], Apostolos Tsakyridis[(1)], Miltiadis Moralis-Pegios[(1)], Christos Pappas[(1)], Manos Kirtas[(1)], Nikolaos Passalis[(1)], David Lazovsky[(2)], Anastasios Tefas[(1)], Nikos Pleros[(1)]

[(1)] Department of Informatics, Center for Interdisciplinary Research & Innovation, Aristotle University of Thessaloniki, Thessaloniki, Greece, email: giamouge@csd.auth.gr
[(2)] Celestial AI, 100 Mathilda Place, Suite 170, Campbell, CA 95008, United States

Abstract *We demonstrate neuromorphic silicon photonic computing that supports fast input/weight update rates together with dynamic precision capabilities, validating experimentally the classification of the IRIS dataset within a two-layer NN with compute speeds up to 50 GHz.*

Introduction

As Moore's law slows down [1], Optical Neural Networks (ONNs) appear as a promising candidate to effectively sustain the tremendous compute growth that current applications demand, due to their high-bandwidth and low-power consumption credentials [2]. Yet, migration to analog optical computing is facing two major challenges towards meeting the performance levels of current digital NN engines: (i) the execution of large NNs with tens of neural layers, which can hardly fit entirely into any hardware platform, together with (ii) the limited bit-precision of analog optical computing engines.

In particular, upgrading ONNs into a general-purpose AI processing platform has to proceed along the paradigm of today's TPU and GPU computational settings, where a limited amount of hardware resources can execute deep NNs with significantly higher dimensions. This can be accomplished only by splitting matrices in smaller tiles and performing tiled matrix multiplication via time division multiplexing (TDM) [3], necessitating the use of neuromorphic photonic solutions that support not only fast input but also fast weight update rates. However, the majority of research in the neuromorphic photonics field has emphasized on NN implementations that support solely static weighting schemes, investing in photonic weight technologies with slow reconfiguration times.

At the same time, the limited bit precision of the constituent high-speed/high bandwidth photonic building blocks [4], bears, inevitably, significant accuracy degradation during the inference process. To this end, in pre-trained networks, analog optical computing can offer tangible benefits over its electronic counterparts, only when operations can be executed at low bit precision [5]. As of today, researchers have made intensive efforts towards mitigating these noise-related challenges. The authors in [6]-[8], proposed a method to tackle the noise induced by the analog photonic hardware via noise-aware pre-trained networks, incorporating various noise sources in the training process. Yet, even though these techniques lead to accuracy improvements, they impose additional complexity and energy trade-offs since the NN need to be retrained in order to be tailored to the employed hardware constraints.

In this paper, we present a neuromorphic photonic processor capable to perform tiled matrix multiplication through TDM and dynamic precision noise-aware inference via the effective reconfiguration of the data rate among the NN's layers. The proposed methods were experimentally applied on the classification of the IRIS dataset [9] via an integrated SiPho chip, executing a total number of 70 MACs over a 2-MAC-supporting neuron and revealing the classification accuracy dependence on the employed data rate. Specifically, the noise analysis of the NN unveiled high noise tolerance on its first layer and a noise sensitive output layer. This study was experimentally validated by carrying out the inference of the constituent neural layers via TDM and recording the accuracy of the NN when the linear operations of the two layers were performed at 2, 16 and 50 Gbaud. The software accuracy of 96.6% was achieved at the experimental inference of the first layer irrespective of the employed compute rate and 93.3%, 86.8% and 68.8% accuracies were obtained when the second layer was implemented at 2, 16 and 50 Gbaud, respectively.

Noise-aware NN inference using TDM

In view of the classification of the IRIS dataset we designed and trained the NN illustrated in Fig. 1 (a). Specifically, the 4 features of the IRIS flowers comprise the inputs of the 4:10:3 fully connected NN. The training was performed employing float-32 single precision variables, achieving a classification accuracy of 96.6%. To

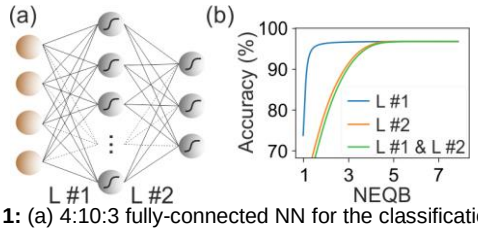

Fig. 1: (a) 4:10:3 fully-connected NN for the classification of the IRIS dataset. (b) Inference quantization bits requirements per neural layer.

quantify the classification accuracy and bit-resolution dependence of each neural layer during inference, we investigate their noise equivalent quantization bit (NEQB) requirements individually [10]. Specifically, we quantize the NN parameters of the examined layer with 1-8 bits. As Fig. 1(b) reveals, approximately a NEQB of 2 is required to achieve the maximum accuracy at the 1st layer, while the respective bit requirements for the 2nd layer was measured to be equal to ~5 bits. This analysis comprises the key for the effective selection of the inference conditions per layer i.e., compute rate and bit precision, leading to an efficient post-training, noise-aware and dynamic-rate NN inference.

On top of the above, since the deployed ONN hardware size cannot follow the NN dimensions, the inference of the NN should be effectively unfolded in time [11]. Figure 2(a) illustrates a microscope photo of the fabricated SiPho processor, that employs electro-absorption modulators (EAMs) for encoding the NN parameters and corresponds to a 2:1 neuron [12]. Hence, the fabricated 2:1 neuron needs to be effectively reused to execute the linear operations of the NN trained for the IRIS dataset (4:10:3). Figure 2 (b) depicts an indicative example of the intra-neuron TDM employed for the compound of 4 inputs x_1-x_4 and their respective weights w_1-w_4, decomposed into 2

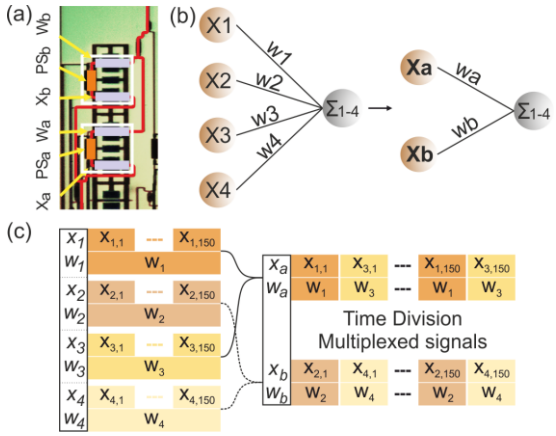

Fig. 2: (a) Microscope photo of the integrated silicon photonic 2-input neuron. The elementary computational cells are encapsulated within the white rectangles. (b) Implementation of a 4–input neuron through a 2:1 neuron (c) Intra-neuron time division multiplexing of both the NN inputs and weight parameters.

sequences x_a,x_b and w_a,w_b, respectively, in order for a 4-input neuron of Layer #1 to fit into the 2:1 photonic neuron [11]. Figure 2(c) illustrates the TDM of the four inputs x_1-x_4, along with their respective weights, w_1-w_4, into the x_a, x_b and w_a, w_b data sequences. More specifically, the multiplexed input data, x_a and x_b are formed as $x_{1,1}x_{3,1}x_{1,2},x_{3,2}\ldots x_{1,N}x_{3,N}$ and $x_{2,1}x_{4,1}x_{2,2},x_{4,2}\ldots x_{2,N}x_{4,N}$ respectively, where $x_{i,j}$ refers to the j^{th} sample ($j \in [1,150]$) of the i^{th} ($i \in [1,4]$) neuron input and its sample number. Their respective weights w_a, w_b were formulated accordingly. Finally, intra-layer TDM was applied towards the implementation of the linear operations of the whole neural layer and subsequently of the entire NN.

Experimental setup and results

Towards the execution of the inference of the IRIS dataset, we established the experimental testbed depicted in Fig. 3. Specifically, a 10 mW continuous wave signal was generated at 1560 nm and injected to the SiPho processor. The signal was then split via a 3 dB splitter into the two identical arms of a Mach-Zehnder Interferometer, that are composed of two EAMs and a thermo-optic heater each [12]. The first were employed for the encoding of the NN input and weight parameters, while the latter safeguards the constructive interference of the weighted inputs [13], since both the inputs and the weights of the NN were forced to be positive during the NN training. A 38 GHz arbitrary waveform generator (AWG) and a DC control plane were employed for the digital to analog conversion of the RF and the DC signals, respectively. The positive linear summation of the weighted inputs was coupled out of the SiPho chip, captured by a 70 GHz photodiode and digitized via a 66 GHz real time oscilloscope. Finally, an NN-library was established to orchestrate the whole inference procedure. More specifically, a series of digital signal processing steps were applied in the transmitter site, that comprised the decoupling and multiplexing of the NN parameters, the compensation of the non-linearities of the electronic-photonic link, the resampling, the pulse shaping and the quantization of the signals before being uploaded to the AWG. Thereafter, in the receiver site, the digitized signals were time recovered, filtered, demultiplexed and downsampled. Finally, the activation functions and the calculation of the NN accuracy were performed in software.

In order to benchmark the noise profile of the ONN in different compute rates, we experimentally characterized the noise standard deviation of the electro-photonic link and correlated it with the achievable NEQB, as

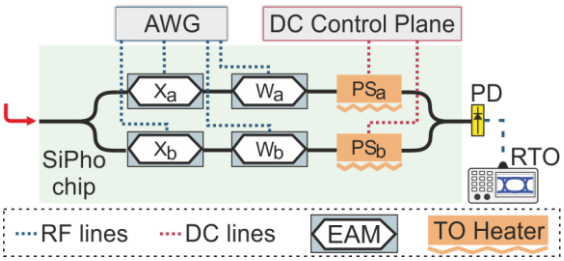

Fig. 3: Experimental testbed for the evaluation of the silicon integrated 2-input neuron.

illustrated in Fig. 4(a). Thereafter, following the TDM and noise aware inference methods described in the previous sections, we classified 150 samples of the IRIS dataset. Specifically, we assessed the experimental achieved accuracy of each of the two neural layers separately and benchmarked it versus their requirements in quantization bits during inference, calculated via the per-layer NEQB analysis. As such, the execution of the first neural layer at a rate as high as 50 Gbaud without degrading the classification accuracy performance was validated, while the output layer proved to be less tolerant to the quantization noise. This is, also, revealed in Fig. 4(b)-(d), where the first 50 samples of output neuron #1 are presented, when the inference of the output layer was performed at 2, 16 and 50 Gbaud, respectively. As expected, the divergence of the experimentally derived curves (blue) from their respective reference (orange) becomes more pronounced as the compute rate increases. In order to quantify this divergence, we plot its probability density function at different operating compute rates, shown in Fig. 4(e), revealing a standard deviation of 0.049, 0.09 and 0.218 for the data rates of 2, 16 and 50 Gbaud, respectively. Hence, depending on the targeted classification accuracy, we enable the selection of the corresponding compute rate for the execution of the output layer. More specifically, Fig. 4(f) depicts the NEQB per layer versus the classification accuracy, with the dashed and solid lines corresponding to the simulated derived curves from the noise aware inference method for the 1st and 2nd layer, respectively. The scatter points represent the experimentally obtained accuracy for the operating compute rates that closely match the simulated accuracy with respect to the NEQB derived from the experimentally calculated noise standard deviation of the operating rates, as shown in Fig. 4(a). Finally, Fig. 4(g) illustrates the experimentally obtained classification accuracies at 2, 16 and 50 Gbaud compute rates, compared with the accuracy derived by the software. The results reveal that the accuracy degradation emerges at the final layer, as has been also identified previously, and becomes even more pronounced as the compute rate increases, with 2G operation yielding 93.3%, 16G 86.6% and 50G 68.8%.

Acknowledgements

This work was supported by the European Commission via H2020 Projects PLASMONIAC (871391), SIPHO-G (101017194) and by the Hellenic Foundation for Research and Innovation (H.F.R.I.) through project DeepLight (4233).

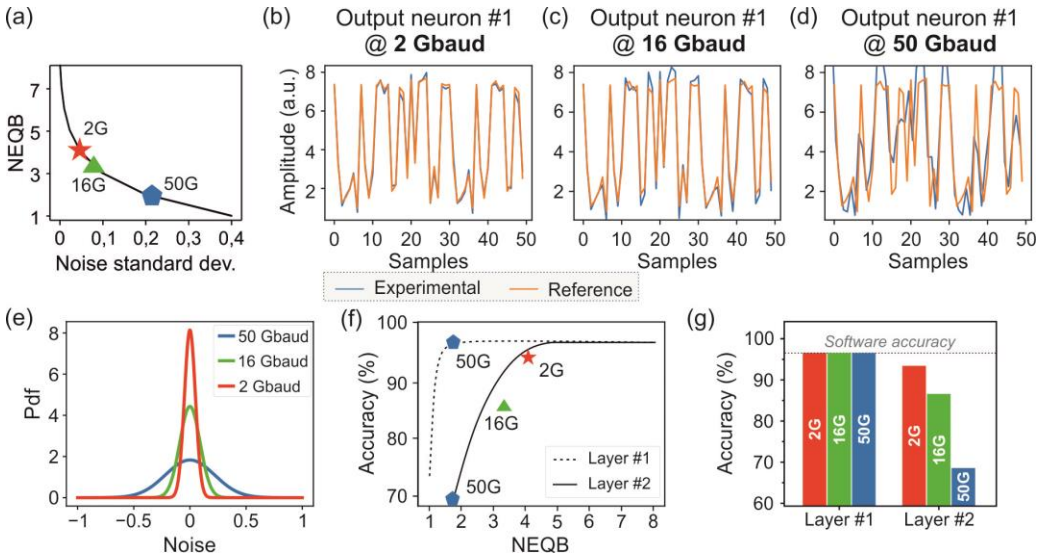

Fig. 4: (a) Opto-electronic link noise variance versus NEQB. Indicative trace of the first 50 samples of the upper output neuron computed at 50 Gbaud at the first layer and (b) at 2, (c) 16 and (d) 50 Gbaud at the second. The software and the experimentally received signals correspond to the orange and the blue curves, respectively. (e) Probability density function of the experimentally obtained traces of the second layer at 2, 16 and 50 Gbaud. (f) NEQB versus the achievable NN accuracy, when the first (dashed line) or the second (solid line) layer's data is quantized with [1,8] bits. (g) Experimentally acquired classification accuracy per neural layer at 2, 16 and 50 Gbaud.

References

[1] T. N. Theis and H. -S. P. Wong, "The End of Moore's Law: A New Beginning for Information Technology," in Computing in Science & Engineering, vol. 19, no. 2, pp. 41-50, Mar.-Apr. 2017, doi: 10.1109/MCSE.2017.29.

[2] A. R. Totović, G. Dabos, , N. Passalis, A. Tefas,, & N. Pleros,(2020). Femtojoule per MAC Neuromorphic Photonics: An Energy and Technology Roadmap. IEEE Journal of Selected Topics in Quantum Electronics, 26, 1-15.

[3] NVIDIA app. Note [Online]. Available: https://docs.nvidia.com/deeplearning/performance/dl-performance-matrix-multiplication/index.html.

[4] S. Garg, J. Lou, A. Jain, M. Nahmias, "Dynamic Precision Analog Computing for Neural Networks", Feb. 2021, [Online]. Available: https://arxiv.org/abs/2102.06365.

[5] M. A. Nahmias, T. F. de Lima, A. N. Tait, H. Peng, B. J. Shastri and P. R. Prucnal, "Photonic Multiply-Accumulate Operations for Neural Networks," in IEEE Journal of Selected Topics in Quantum Electronics, vol. 26, no. 1, pp. 1-18, Jan.-Feb. 2020, Art no. 7701518, doi: 10.1109/JSTQE.2019.2941485.

[6] M. Moralis-Pegios, G. Mourgias-Alexandris, A. Tsakyridis, G. Giamougiannis, A. Totovic, G. Dabos, N. Passalis, M. Kirtas, A. Tefas, N. Pleros, "Neuromorphic Silicon Photonics and Hardware-aware Deep Learning for High-Speed Inference," in Journal of Lightwave Technology, doi: 10.1109/JLT.2022.3171831.

[7] G. Mourgias-Alexandris, M. Moralis-Pegios, A. Tsakyridis, N. Passalis, M. Kirtas, A. Tefas, T. Rutirawut, F. Y. Gardes, and N. Pleros, "Channel response-aware photonic neural network accelerators for high-speed inference through bandwidth-limited optics," Opt. Express 30, 10664-10671 (2022)

[8] E. Paolini, L. De Marinis, M. Cococcioni, L. Valcarenghi, L. Maggiani, N. Andriolli, "Photonic-aware neural networks". Neural Comput & Applic (2022). https://doi.org/10.1007/s00521-022-07243-z

[9] [Online]: https://archive.ics.uci.edu/ml/datasets/iris

[10] C. Pearson, "High-speed, analog-to-digital converter basics," Texas Instruments, Dallas, TX, USA, App. Rep. SLAA510, Jan. 2011. [Online]. Available: http://www.ti.com/lit/an/slaa510/slaa510.pdf

[11] A. Tsakyridis, G. Giamougiannis, G. Mourgias-Alexandris, A. Totovic, G. Dabos, N. Passalis, M. Kirtas, A. Tefas, M. Moralis-Pegios, N. Pleros, "Silicon Photonic Neuromorphic Computing with 16 GHz Input Data and Weight Update Line Rates", CLEO 2022.

[12] G. Giamougiannis, A. Tsakyridis, G. Mourgias-Alexandris, M. Moralis-Pegios, A. Totovic, G. Dabos, N. Passalis, M. Kirtas, N. Bamiedakis, A. Tefas, D. Lazovsky, N. Pleros, "Silicon-integrated coherent neurons with 32GMAC/sec/axon compute line-rates using EAM-based input and weighting cells" in ECOC (2021).

[13] G. Mourgias-Alexandris, A. Totovic, A. Tsakyridis, N. Passalis, K. Vyrsokinos, A. Tefas, N. Pleros "Neuromorphic Photonics With Coherent Linear Neurons Using Dual-IQ Modulation Cells," in Journal of Lightwave Technology, vol. 38, no. 4, pp. 811-819, 15 Feb.15, 2020, doi: 10.1109/JLT.2019.2949133.