Optics EXPRESS

Channel response-aware photonic neural network accelerators for high-speed inference through bandwidth-limited optics

G. Mourgias-Alexandris,[1,2,*] M. Moralis-Pegios,[1,2] A. Tsakyridis,[1,2] N. Passalis,[1,2] M. Kirtas,[1,2] A. Tefas,[1,2] T. Rutirawut,[3] F. Y. Gardes,[3] and N. Pleros[1,2]

[1] *Department of Informatics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece*
[2] *Center for Interdisciplinary Research and Innovation (CIRI-AUTH), Balkan Center, Buildings A & B, Thessaloniki, 10th km Thessaloniki-Thermi Rd, P.O. Box 8318, GR 57001, Greece*
[3] *Optoelectronics Research Centre, University of Southampton, Southampton SO17 1BJ, UK*
[*] *mourgias@csd.auth.gr*

Abstract: Photonic neural network accelerators (PNNAs) have been lately brought into the spotlight as a new class of custom hardware that can leverage the maturity of photonic integration towards addressing the low-energy and computational power requirements of deep learning (DL) workloads. Transferring, however, the high-speed credentials of photonic circuitry into analogue neuromorphic computing necessitates a new set of DL training methods aligned along certain analogue photonic hardware characteristics. Herein, we present a novel channel response-aware (CRA) DL architecture that can address the implementation challenges of high-speed compute rates on bandwidth-limited photonic devices by incorporating their frequency response into the training procedure. The proposed architecture was validated both through software and experimentally by implementing the output layer of a neural network (NN) that classifies images of the MNIST dataset on an integrated SiPho coherent linear neuron (COLN) with a 3dB channel bandwidth of 7 GHz. A comparative analysis between the baseline and CRA model at 20, 25 and 32GMAC/sec/axon revealed respective experimental accuracies of 98.5%, 97.3% and 92.1% for the CRA model, outperforming the baseline model by 7.9%, 12.3% and 15.6%, respectively.

1. Introduction

Deep Learning-based NNs have had a profound effect on computer science during the last decade, revolutionizing a number of applications such as finance, health, and telecommunications [1]. The constantly growing requirements of modern DL workloads, however, has fueled the deployment of custom hardware towards accelerating performance, calling for Multiply-And-Accumulate (MAC) operations within a low-energy and high-bandwidth envelope. In this context, PNNAs have emerged as a promising technological candidate that can leverage the growing maturity of photonic integration with the well-established speed and power consumption properties of optical circuits. This roadmap builds upon a solid perspective for orders of magnitude improvements in both the computational rates and the energy efficiency compared to their electronic counterparts [2–4]. However, PNNA implementations reported so far, based either on Wavelength Division Multiplexing (WDM) or coherent layouts [5–14], could hardly scale their compute rates beyond the kHz or MHz range, with the performance of 10GMAC/sec/axon being recently accomplished both in incoherent [12] and coherent [14] PNNAs.

Inter-Symbol Interference (ISI) and thermal noise comprise the two main factors for trading off compute rate speeds for accuracy performance. Efforts for mitigating their impact have mainly emphasized in quantifying their presence and then treating them as an inherent part of the underlying hardware, adapting DL training models and algorithms over the specific characteristics

#452803 https://doi.org/10.1364/OE.452803
Journal © 2022 Received 3 Jan 2022; revised 3 Feb 2022; accepted 4 Feb 2022; published 15 Mar 2022

of the analogue photonic platform [15–18]. In this context, the quantization noise originating from the Digital-to-Analog and Analog-to-Digital Converters (DAC and ADC) was studied in [19], validating the energy gains of specially quantized NN models, in limited bit-resolution use cases. Moreover, the effects of non-deterministic noise sources that were approximated via Additive Gaussian Noise Sources (AWGN), including laser Relative Intensity Noise (RIN), Johnson shot-noise and uniform quantization noise, were also extensively studied in [16–18], concluding to noise-resilient DL training models that were recently also verified experimentally [20,21]. However, ISI contributions continue to form an unwanted inference accuracy loss factor when the underlying hardware has non-flat frequency response, comprising a performance limiting term that is typically completely ignored by DL models over digital hardware platforms.

In this paper, we present and experimentally demonstrate for the first time, to the best of our knowledge, a new CRA training model that integrates the frequency response of the constituent photonic channel in the training of a PNNA. The proposed method extends our preliminary results of [22] by increasing the experimentally demonstrated compute rate performance to 32GMAC/sec/axon and analyzing through simulations the network accuracy versus the compute rate-bandwidth ratio when used as the MNIST classification layer over the whole dataset of MNIST digits. The experimental validation has been carried out on an integrated SiPho Coherent Linear Neuron (COLN), with the CRA training model incorporating the fan-in Mach-Zehnder Modulator (MZM) transfer function into the NN axon response. A comparative analysis between the baseline and CRA models at 20, 25 and 32GMAC/sec/axon compute rates reveal respective experimental accuracies of 98.5%, 97.3% and 92.1% for the CRA model and 90.6%, 85% 76.5% for the baseline model. With the 3-dB bandwidth of the SiPho COLN being only 7GHz and mainly dictated by the fan-in MZM, the performance of the CRA model outperforms the baseline model by 7.9%, 12.3% and 15.6% at 20, 25 and 32GMAC/sec/axon compute rates, respectively, leading to a speed versus bandwidth ratio of >4.

2. Concept and NN architecture

Figure 1(a) depicts a generic schematic layout of an N-fan-in coherent linear neuron, following the architectural approach presented in [11] and experimentally demonstrated in [14]. The layout comprises a single bias branch, that safeguards negative weight imprinting, and an N number of axons, each consisting of an amplitude modulator for imprinting the input data x_{i} followed by a phase shifter s_i for imprinting the weight sign information cascaded with a weight amplitude $|w_{\mathrm{i}}|$ modulator. As such, the weighted inputs i.e $x_{\mathrm{i}}*w_{\mathrm{i}}$, emerging at each axon output, are combined at the neuron's output coupler with the bias branch to yield the total weighted sum $\sum_i^N XiWi$. All three data imprinting building blocks are driven by respective electrical driving signals, assuming static values for the weight values i.e w_{i} and dynamic values for the input data x_{i}. A breakdown of the most significant noise sources impacting the optical signal as it traverses through a single neuron axon is illustrated in Fig. 1(b). Optical noise originating from the laser source, including RIN and phase noise, is denoted as n_{laser}, while the noise contributions stemming from the input data amplitude modulator, the phase shifter and the weight amplitude modulator are denoted as n_{x}, n_{s} and n_{w}, respectively. Finally, f_{x} corresponds to the frequency response of the input data modulator, that is driven by a high-speed RF signal and introduces ISI in the signal's path.

In order to investigate the effect of the ISI originating from the COLN's channel response and its impact on a PNNA, we designed a NN, depicted in Fig. 1(c). The NN was trained for classifying images of the MNIST dataset and incorporated a specially designed software building block that allows the inclusion of the COLN's modulators channel frequency response, during both the training and inference phase. The designed NN relies exclusively on fully connected feed-forward neurons and comprises the input layer followed by 2 hidden layers and the photonic output layer. The input data stream originating from hidden layer #2 is converted to the frequency domain via Real Fast Fourier Transformation (RFFT), and then multiplied with an

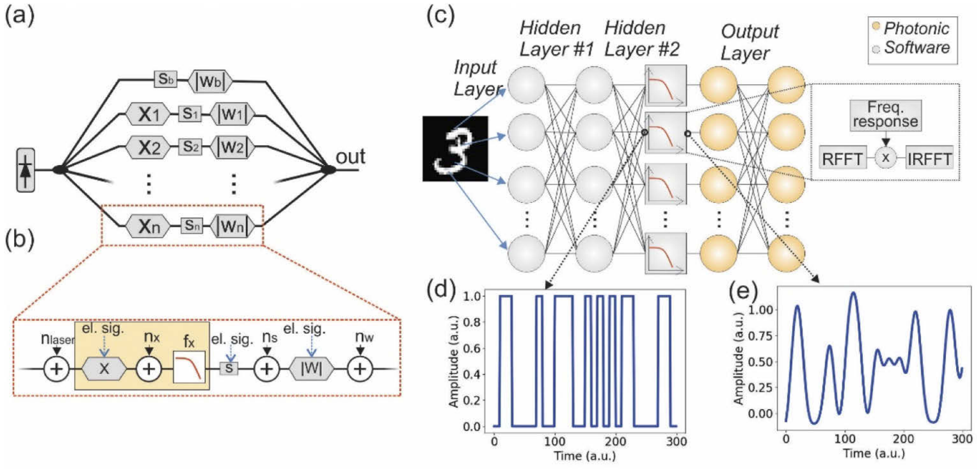

Fig. 1. (a) Typical layout of an N-fan-in coherent linear neuron (b) Detailed schematic of a single axon including the major noise sources, (c) Designed architecture for the investigation of ISI effect in PNNs, (d) indicative pseudo-random time trace in the PNNA input (e) time-trace at the output of the neuron impacted by the limited transfer function of the photonic building blocks.

arbitrary channel response in the frequency domain. The resulting signal is then converted back to the time domain through an Inverse Real Fast Fourier Transformation (IRFFT). In this way, the proposed photonic NN architecture allows for the precise incorporation of any real-valued channel response into the NN, enabling in this way CRA training and inference for PNNAs with amplitude-encoded output. The effect of this low-pass channel response is visualized in Fig. 1(d) and (e), with Fig. 1(d) showing a time trace of a pseudo-random signal representing the NN output, while Fig. 1(e) illustrates the waveform after traversing a low pass filter stage, that emulates the frequency response of the x_i input imprinting modulators. As can be observed, the signal in Fig. 1(e) is distorted by the deterministic noise that attenuates more significantly the short duration pulses, introducing low pass filtering within the PNNA and as such significantly reducing the neuron's output Signal-to-Noise-Ratio (SNR) and as such the accuracy of the NN-based classifier.

3. Experimental setup & results

In order to assess and quantify the performance of the proposed DL model, we implemented the output layer of the NN depicted in Fig. 1(c), using an integrated 4-input SiPho COLN [14] designed using PDK-ready components from the CORNERSTONE Multi-Project Wafer (MPW) service. Two models were assessed both in software and hardware: i) The BaseLine (BL) model, with the NN trained without accounting for any bandwidth limitations, following a more conventional training procedure where the channel response of the photonic hardware is not taken into account. ii) The CRA model, where we incorporated the experimentally obtained frequency response of the SiPho COLN, in the NN training procedure, utilizing the specially designed block described in Section 2. The normalized magnitude response of the derived SiPho COLN transfer function is depicted in the inset of Fig. 2(a), revealing a low-pass shaped function with a 3dB bandwidth of 7 GHz. The NN training was performed at 20, 25 and 32GMAC/sec/axon compute rates, resulting in three CRA and three baseline models. For all scenarios the training lasted for 40 epochs, with a batch size of 100, while the Adam optimizer was employed with a learning rate equal to 0.001.

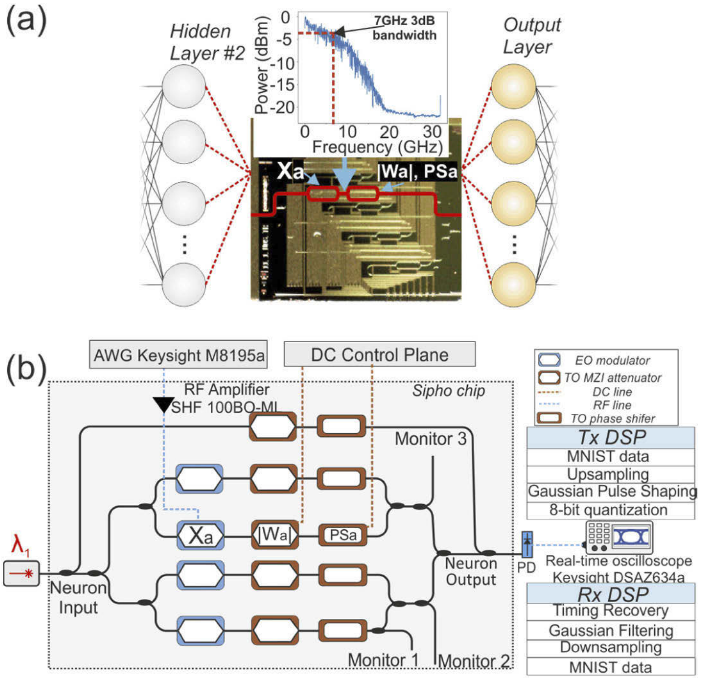

Fig. 2. (a) Schematic illustration of the last two layers of the proposed PNN, with the output layer realized through a silicon photonic COLN, (b) Experimental setup used for the experimental comparison of the CRA and baseline approaches.

Figure 2(b) illustrates the experimental setup used for the assessment of all the afore-mentioned NN models, with the output layer of the NN deployed on the SiPho COLN. The x_i values of the NNs output layer were imprinted sequentially on a single axon of the SiPho COLN utilizing the same silicon MZM, while the weighting was implemented offline through software. Light from a laser beam at λ_1=1554.55nm, was injected into the SiPho through a Photonic Design Kit (PDK)-ready TE grating coupler. An EO-MZM x_a biased on its quadrature point was used to convert the corresponding NN data from the electrical to the optical domain, while the thermo-optic elements $|w_{\ a}|$ and PS_a were configured to produce a weighting value equal to 1. Before reaching the COLN, the NN data were upsampled to 3, 2.4 and 1.875 samples per symbol (sps) in order to realize the 20, 25 and 32GMAC/sec compute rates. Furthermore, a Gaussian filter with a band factor σ=0.7 has been in order to realize a signal with low-pass response, achieving in this way the least possible inter-symbol interference. Then, 8-bit quantization has been performed before the signal being uploaded to Keysight's M8195a Arbitrary Waveform Generator (AWG) operating at 60GSa/s. Each MNIST grayscale image was matched perfectly with the 8-bit resolution of the AWG since the grayscale has 256 (i.e. 2^8) discrete values. Furthermore, the DAC outputs were operated at the maximum V_{pp} (i.e., =1V) to achieve the highest possible Signal-to-Noise Ratio, with each grayscale level corresponding to a voltage difference of ~3.9mV$_{pp}$. The AWG's differential outputs were then amplified by two SHF100BO-ML RF amplifiers to drive the push-pull EO-MZM with ~3Vpp. Consecutively, the resulting optical signal was converted back to the electrical domain via a PIN photodetector with 50GHz 3dB bandwidth. The output of the PD was captured by a Keysight DSOZ632a Real Time Oscilloscope (RTO) with 80GSa/s and 33GHz bandwidth. Time-synchronization and Gaussian filtering was performed on the captured waveform before being downsampled to 1sps and forwarded back to the NN. The same procedure was followed during the deployment of each BL and CRA model at 20, 25 and 32GMAC/sec, respectively.

Figures 3(a) and (b) illustrate the experimentally obtained (red curves) versus the NN expected (blue curves) time traces for the BL and CRA models, when the integrated SiPho COLN implemented the output layer of the respective NN at 25 GMAC/sec/axon. As can be observed, the BL model derived trace, diverges significantly from the NN expected one, as it is significantly affected by ISI originating from the limited 7 GHz bandwidth of the COLN. Deployment of the CRA model achieves a much better matching of the experimental and expected time traces, as the specific NN training effectively cancels out the bandwidth limitations and related ISI.

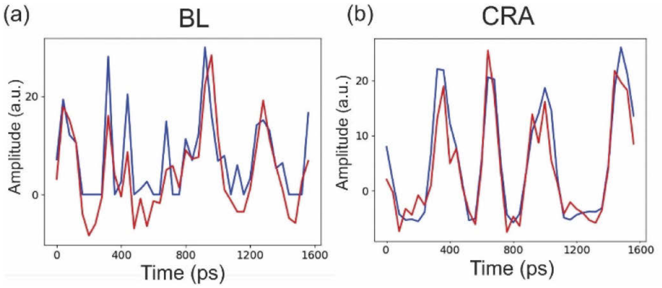

Fig. 3. Time traces of the expected (blue curve) and obtained (red curve) NN data at 25GMAC/sec/axon of the (a) BL model at (b) CRA model.

The difference between the expected and the experimentally obtained signals was also quantified via the accuracy measurements and respective confusion matrixes depicted in Figs. 4(a)-(f). Figure 4(a) illustrates the experimentally obtained accuracy of the BL model at 20GMAC/sec/axon, where the accuracy for each MNIST digit spanned in the range of [78.4%- 99.4%]. The accuracy for the corresponding CRA model is depicted in Fig. 4(b), revealing improved accuracy within the range [94.7%-99.5%]. The same set of results for both models at 25 and 32GMAC/sec/axon compute rates are illustrated in Figs. 4(c)-(f), where the accuracy improvement due to the CRA model is much higher compared to 20GMAC/sec/axon case. The average experimentally obtained accuracy for all MNIST digits is depicted in Fig. 5 for each BL and CRA model, as well as the simulation results for the BL and the CRA model spanning from 7 to 42GMAC/sec/axon using a step of 7GMAC/sec/axon that is equivalent to a step of 1 for the compute rate versus bandwidth. The software deployment of BL model revealed a classification accuracy equal to 98.9%, while the simulated BL performance in the range between 7-42GMAC/axon/sec decreased from 98.8% to 63%. The experimental validation of this model revealed an accuracy of 90.6%, 85% and 76.5% at 20, 25 and 32GMAC/sec/axon. The software deployment of CRA model in the range of (7, 42) GMAC/sec/axon revealed accuracies between 99% and 84%, respectively. The accuracy for the three experimentally validated compute rates was 98.6%, 98% and 93.2% at 20, 25 and 32GMAC/sec/axon, respectively, while the experimental evaluation of the models followed closely the software performance revealing 98.5%, 97.3% and 92.1% classification accuracy at 20, 25 and 32GMAC/sec/axon, respectively.

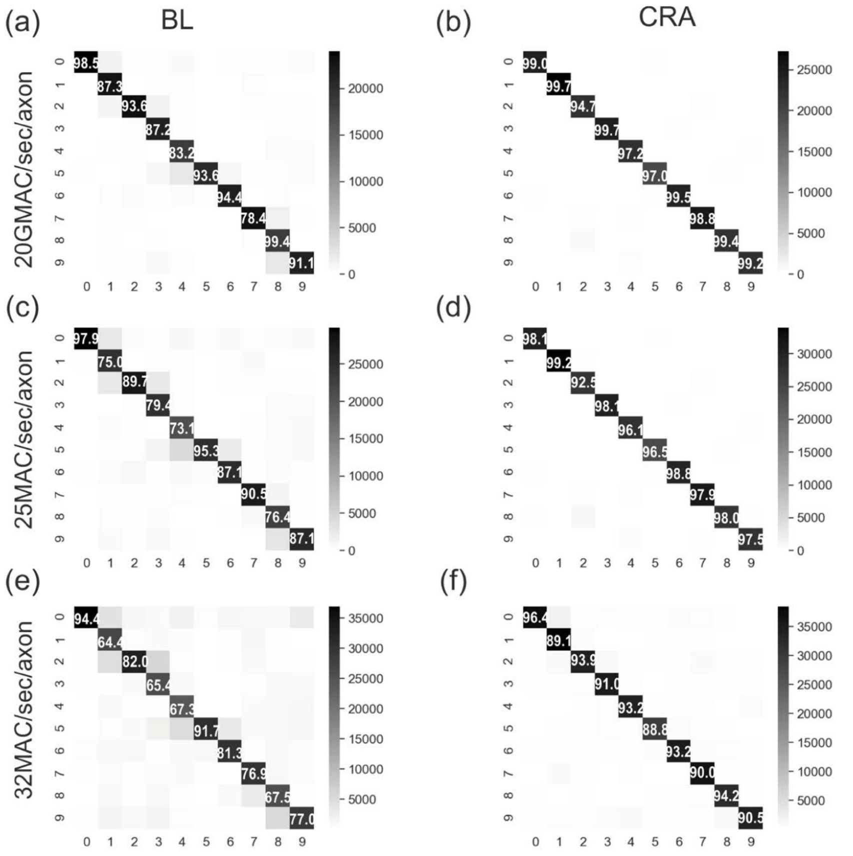

Fig. 4. Confusion matrices representing the experimentally obtained classification accuracy of each digit from MNIST dataset for the Baseline (BL) and CRA model: (a), (b) 20, (c), (d) 25 and (e), (f) 32GMAC/sec/axon, respectively.

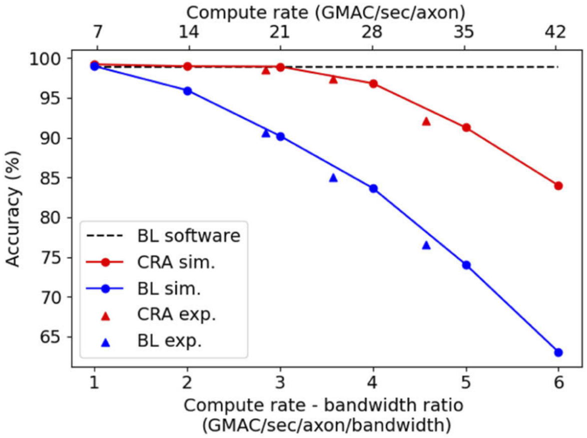

Fig. 5. The average software and experimentally obtained accuracy of each compute rate for the baseline and CRA models, respectively. The simulations for both models are depicted with solid lines in the range of (7, 42) GMAC/sec/axon compute rates.

4. Conclusion

We presented a novel CRA NN architecture, that incorporates the frequency response of the underlying photonic components into the training procedure. The performance of the proposed scheme was experimentally assessed by implementing the output layer of a NN trained for classifying digits of the MNIST dataset, on a 4-input integrated COLN [14], with a 3dB bandwidth of 7 GHz. Experimental accuracies of 98.5%, 97.3% and 92.1% were reported, for compute rates of 20, 25 and 32GMAC/sec/axon, respectively. On top of that, the BL and CRA models were benchmarked via simulations within the range of (7, 42) GMAC/sec/axon, revealing classification accuracies between (98.8%, 63%) and (99%, 84%), respectively. The CRA models outperformed the BL models for every MNIST digit and compute rate, while at the same time reduced the accuracy discrepancy between software simulation and experimental deployment revealing the matching of the trained NN to the specific photonic components response. Furthermore, this work validated the performance credential of DL-inspired training frameworks to compensate ISI distortions originating from the photonic circuitry, enabling ultra-fast PNNA engines based on low-bandwidth and low-cost optics without using any digital signal processing technique. A similar architecture could be used to tolerate different types of deterministic noise such as deterministic jitter. Finally, the synergy of channel-response aware and noise-aware architectures could provide a powerful training framework towards reliable and low-cost photonic neural network accelerators.

Funding. Hellenic Foundation for Research and Innovation (4233, DeepLight); European Commission (871391, PLASMONIAC).

Disclosures. The authors declare that there are no conflicts of interest related to this article.

Data availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

References

1. Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," Nature **521**(7553), 436–444 (2015).
2. A. R. Totovic, G. Dabos, N. Passalis, A. Tefas, and N. Pleros, "Femtojoule per MAC Neuromorphic Photonics: An Energy and Technology Roadmap," IEEE J. Sel. Top. Quantum Electron. **26**(5), 1–15 (2020).
3. M. A. Nahmias, T. F. De Lima, A. N. Tait, H. T. Peng, B. J. Shastri, and P. R. Prucnal, "Photonic Multiply-Accumulate Operations for Neural Networks," IEEE J. Sel. Top. Quantum Electron. **26**(1), 1–18 (2020).
4. R. Stabile, G. Dabos, C. Vagionas, B. Shi, N. Calabretta, and N. Pleros, "Neuromorphic photonics: 2D or not 2D?" J. Appl. Phys. **129**(20), 200901 (2021).
5. H. Peng, T. F. de Lima, M. A. Nahmias, A. N. Tait, B. J. Shastri, and P. R. Prucnal, "Autaptic Circuits of Integrated Laser Neurons," in *Conference on Lasers and Electro-Optics* (OSA, 2019), Vol. 1, p. SM3N.3.
6. A. N. Tait, T. F. De Lima, M. A. Nahmias, H. B. Miller, H. Peng, B. J. Shastri, and P. R. Prucnal, "Silicon Photonic Modulator Neuron," Phys. Rev. Appl. **11**(6), 064043 (2019).
7. A. N. Tait, T. F. De Lima, E. Zhou, A. X. Wu, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Neuromorphic photonic networks using silicon photonic weight banks," Sci. Rep. **7**(1), 1–10 (2017).
8. G. Mourgias-Alexandris, N. Passalis, G. Dabos, A. Totovic, A. Tefas, and N. Pleros, "A Photonic Recurrent Neuron for Time-Series Classification," J. Lightwave Technol. **39**(5), 1340–1347 (2021).
9. H. T. Peng, M. A. Nahmias, T. F. De Lima, A. N. Tait, B. J. Shastri, and P. R. Prucnal, "Neuromorphic Photonic Integrated Circuits," IEEE J. Sel. Top. Quantum Electron. **24**(6), 1–15 (2018).
10. Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and M. Soljačić, "Deep learning with coherent nanophotonic circuits," Nat. Photonics **11**(7), 441–446 (2017).
11. G. Mourgias-Alexandris, A. Totovic, A. Tsakyridis, N. Passalis, K. Vyrsokinos, A. Tefas, and N. Pleros, "Neuromorphic Photonics With Coherent Linear Neurons Using Dual-IQ Modulation Cells," J. Lightwave Technol. **38**(4), 811–819 (2020).
12. X. Xu, M. Tan, B. Corcoran, J. Wu, A. Boes, T. G. Nguyen, S. T. Chu, B. E. Little, D. G. Hicks, R. Morandotti, A. Mitchell, and D. J. Moss, "11 TOPS photonic convolutional accelerator for optical neural networks," Nature **589**(7840), 44–51 (2021).
13. J. Feldmann, N. Youngblood, M. Karpov, H. Gehring, X. Li, M. Stappers, M. Le Gallo, X. Fu, A. Lukashchuk, A. S. Raja, J. Liu, C. D. Wright, A. Sebastian, T. J. Kippenberg, W. H. P. Pernice, and H. Bhaskaran, "Parallel convolutional processing using an integrated photonic tensor core," Nature **589**(7840), 52–58 (2021).

14. G. Mourgias-Alexandris, M. Moralis-Pegios, S. Simos, G. Dabos, N. Passalis, M. Kirtas, T. Rutirawut, F. Y. Gardes, A. Tefas, and N. Pleros, "A Silicon Photonic Coherent Neuron with 10GMAC/sec processing line-rate," *Optical Fiber Communication Conference (OFC) 2021*, paper Tu5H.1.
15. S. Garg, A. Jain, J. Lou, and M. Nahmias, "Confounding Tradeoffs for Neural Network Quantization," arXiv:2102.06366 (2021).
16. N. Passalis, G. Mourgias-Alexandris, A. Tsakyridis, N. Pleros, and A. Tefas, "Variance Preserving Initialization for Training Deep Neuromorphic Photonic Networks with Sinusoidal Activations," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2019), pp. 1483–1487.
17. N. Passalis, M. Kirtas, G. Mourgias-Alexandris, G. Dabos, N. Pleros, and A. Tefas, "Training noise-resilient recurrent photonic networks for financial time series analysis," *Eur. Signal Process. Conf.*, 1556–1560 (2021).
18. N. Passalis, G. Mourgias-Alexandris, A. Tsakyridis, N. Pleros, and A. Tefas, "Training Deep Photonic Convolutional Neural Networks With Sinusoidal Activations," IEEE Trans. Emerg. Top. Comput. Intell **5**, 384–393 (2021).
19. S. Garg, J. Lou, A. Jain, and M. Nahmias, "Dynamic Precision Analog Computing for Neural Networks," 1–12 arXiv:2102.06365 (2021).
20. G. Mourgias-Alexandris, M. Moralis-Pegios, A. Tsakyridis, S. Simos, G. Dabos, A. Totovic, N. Passalis, M. Kirtas, T. Rutirawut, F. Y. Gardes, A. Tefas, and N. Pleros, "Noise-resilient and high-speed deep learning with coherent silicon photonics" (submitted to Nature Communications).
21. M. Moralis-Pegios, Angelina Totovic, Apostolos Tsakyridis, George Giamougiannis, George Mourgias-Alexandris, George Dabos, Nikolaos Passalis, Manos Kirtas, Anastasios Tefas, and Nikos Pleros, "Photonic Neuromorphic Computing: Architectures, Technologies, and Training Models," *will appear in proceedings of Optical Fiber Comm. Conf. (OFC) 2022*, San Diego, CA, USA, March 2022.
22. G. Mourgias-Alexandris, A. Tsakyridis, N. Passalis, M. Kirtas, A. Tefas, T. Rutirawut, F. Y. Gardes, N. Pleros, and M. Moralis-Pegios, "25GMAC/sec/axon photonic neural networks with 7 GHz bandwidth optics through channel response-aware training," in *2021 European Conference on Optical Communication (ECOC)* (IEEE, 2021), Vol. 1, pp. 1–4.