25GMAC/sec/axon photonic neural networks with 7GHz bandwidth optics through channel response-aware training

George Mourgias-Alexandris [(1)], Apostolos Tsakyridis [(1)], Nikolaos Passalis [(1)], Manos Kirtas [(1)], Anastasios Tefas[(1)] , Teerapat Rutirawut [(2)], Frederic Y. Gardes[(2)] , Nikos Pleros[(1)] , Miltiadis Moralis-Pegios[(1)]

[(1)] Department of Informatics, Aristotle University of Thessaloniki, 54124, Thessaloniki, Greece, mourgias@csd.auth.gr
[(2)] Optoelectronics Research Centre, University of Southampton, Southampton, SO17 1BJ, UK

Abstract We present a channel response-aware Photonic Neural Network (PNN) and demonstrate experimentally its resilience in Inter-Symbol Interference (ISI) when implemented in an integrated neuron. The trained PNN model performs at 25GMAC/sec/axon using only 7GHz-bandwidth photonic axons with 97.37% accuracy in the MNIST dataset.

Introduction

The relentless growth of machine learning workloads has shifted significant research attention in the development of high-bandwidth and energy efficient computing accelerators. Neuromorphic photonics aim to transfer the energy efficiency, high-bandwidth and density credentials of silicon photonics, into the Deep Learning (DL) domain, heralding orders of magnitude higher computational line rates and energy efficiency compared to their electronic counterparts[1], [2]. Previous implementations of Photonic Neural Networks (PNN), based either on WDM or coherent layouts [3]–[12], were however limited in their computational rates, usually demonstrated in the kHz or MHz scale, while only recently the barrier of 10Gbaud was breached[10], [12].

Deployment of PNN in higher computational rates, necessitates the development of deep learning training models that can take into account the physical properties of the employed photonic components and compensate for their non-ideal performance. In this context, previous research on DL models specifically trained for PNN, investigated the effect of deterministic noise originating from signal quantization of DACs and ADCs, validating the robustness of specially trained NNs in quantization limited use cases[13]. Moreover, the effect of non-deterministic noise sources, usually manifested in PNNs in the form of Additive Gaussian Noise Sources (AWGN), was also studied in[14]–[16], revealing that specifically trained DL models can maintain their high accuracy credentials even in highly-noise implementations. However, another significant contributor that should be considered when designing photonic DL models, for high-speed PNN implementations, is the channel response of the employed photonic components. Given the low-pass or non-linear response of the majority of currently deployed SiPho components, ISI is expected to significantly affect the PNN's performance, when targeting high operating rates, in non-specifically trained DL models.

Herein, we present and experimentally demonstrate for the first time, to the best of our knowledge, a new Neural Network (NN) architecture that allows inclusion of the channel response of the photonic components in the training of a PNN. The proposed method was validated both in software and experimentally, on a NN trained for classifying images of the MNIST dataset, with its output layer implemented through an integrated SiPho Coherent Linear Neuron (COLN). The channel response of the modulator used both for the training and the inference stage had a 3dB bandwidth equal to 7 GHz, while a comparison of the channel response-aware and baseline DL models at 20 and 25Gbaud revealed accuracies of 98.51% & 97.37% versus 90.6% & and 85.07% respectively.

Concept and NN implementation

Figure 1 (a) depicts a typical layout of an N-fan-in coherent linear neuron, following the architecture proposed in [9] and experimentally demonstrated in [12]. The architectural layout, is

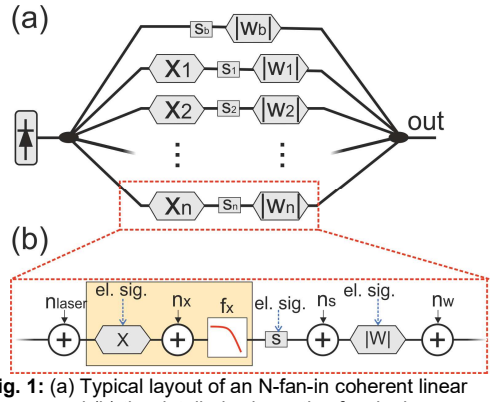

Fig. 1: (a) Typical layout of an N-fan-in coherent linear neuron and (b) the detailed schematic of a single axon.

978-1-6654-3868-1/21/$31.00 ©2021 IEEE

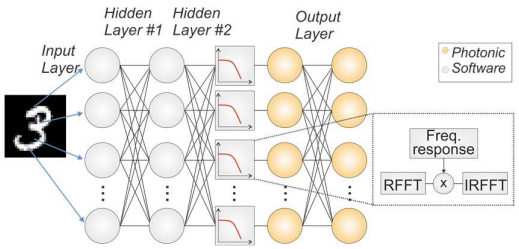

Fig. 2: Designed architecture for the investigation of ISI effect in PNNs.

composed of a single bias branch, that effectively safeguards negative weight imprinting, and an *N* number of axons, each consisting of an amplitude modulator for generating the Input data X_i followed by a phase shifter S_i for providing the weight sign information cascaded with a weight amplitude $|W_i|$ modulator. As such, the basic weighted inputs i.e $X_i * W_i$, 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 Wi and dynamic values for the input data Xi.

A detailed breakdown of the noise sources impacting the optical signal as it traverses a single neuron axon is illustrated in Fig.1 (b). The noise that originates from the laser source is denoted as n_{laser}, while the noises coming from the 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 channel response of the input data modulator, that is driven by a high-speed RF signal and introduces ISI in the signal's path. It should be noted, that the focus of this work has been the study of this deterministic limited frequency response originating noise, with previous works dealing mainly with the remaining noise contributions that can be, without loss of generality, simulated through AWGN[14]–[16].

In order to investigate the effect of the ISI originating from the data input modulator's frequency response and its impact on a PNN, we designed a NN, depicted in Fig. 2. 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 modulators channel response, during both the training and inference stage. 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 arbitrary channel response. 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 channel response of the input data modulator into the NN, enabling in this way channel response-aware training and inference.

Experimental setup & results

The performance of our proposed channel response-aware scheme was assessed through implementing in the PyTorch framework, the aforementioned NN and realizing the functionality of its final output layer on a SiPho COLN, previously demonstrated in[12].

Two different models were evaluated both in software and experimentally: i) The baseline model, were the NN was trained using a flat channel response, resembling the case of a typically trained NN, that does not take into account the channel response of the photonic implementation. ii)The channel response-aware model, that incorporates the experimental derived transfer function of the photonic implementation in the training phase, using our specially designed block described in the previous section. In these experiments, we utilized the transfer function of the input data modulator of the COLN at 20 and 25Gbaud, depicted in the inset of Fig.3(a) which had a 3dB bandwidth of approximately 7 GHz in both cases.

The NNs in both scenarios were trained for 40 epochs, with a batch size of 100, while the Adam optimizer was used to optimize the weights of the

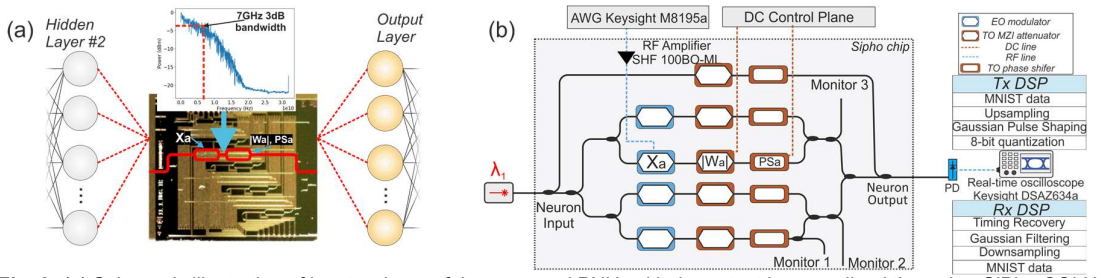

Fig. 3: (a) Schematic illustration of last two layer of the proposed PNN, with the output layer realized through a SiPho COLN (b) Experimental setup used for the assessment of the channel-response aware scheme.

neurons with a learning rate equal to 0.001.

The experimental setup utilized for validating the performance of the NN during both scenarios, when implementing the last output later on the integrated COLN is depicted in Fig. 3(b)[12]. A single axon of the photonic neuron was utilized to sequentially imprint in ascending order the Xi data originating from the 2nd Hidden layer of the designed NN, while weighting was implemented offline through software. A light beam at λ_1=1554.55nm was injected to the SiPho chip via a TE grating coupler. An EO-MZM X_a was used to optically imprint the corresponding NN data, while the $|w_a|$ and PS_a were configured to produce a weighting value equal to 1. In order to interface the NN data to the integrated coherent neuron, their respective waveforms were upsampled from 1 to 3 and 2.4 samples per symbol (sps), corresponding to operational data-rates of 20 and 25Gbaud and were then filtered by a Gaussian filter with σ=0.7. The resulting signals were quantized with 8-bit resolution before being uploaded to Keysight's M8195a Arbitrary Waveform Generator (AWG) operating at 60GSa/s. The output signal and its differential copy originating from the AWG were then forwarded to two SHF100BO-ML RF amplifiers to drive the push-pull EO-MZM with approximately 3Vpp. The optical signal emerging from the SiPho chip was converted to the electrical domain by the means of a PIN photodetector with 50GHz 3dB bandwidth and was subsequently captured by a Keysight DSAZ634a Real Time Oscilloscope (RTO) with 80GSa/s and 33GHz bandwidth. The received signal was time-synchronized with the expected signal, and was then filtered with a Gaussian filter before being downsampled to 1sps and forwarded back to NN. The same procedure was followed for the deployment of NN data at 20 and 25Gbaud, for both the baseline and channel response-aware models.

Figures 4 (a)-(d) depict the experimentally obtained (red curves) versus the NN expected (blue curves) time traces for both the baseline and channel response-aware models, at 20 and 25Gbaud, respectively. As it can be observed, from the captured time traces, the divergence of the received versus expected signals, increases significantly for the baseline model, as the baud rate increases, while a very good matching is achieved in the channel-aware scheme for both cases. This observation is quantized in the reported accuracies values illustrated in Fig.4 (e). The baseline model achieves experimental accuracies of 90.6% and 85.07% on the MNIST classification task at 20 and 25Gbaud, i.e a degradation of 8.3% and 13.83%, versus the baseline software implementation. The software derived accuracies for the channel response-aware scheme reached 98.6% and 98%, a degradation of only 0.3% and 0.9%, respectively. This superior performance of the channel-response aware scheme, was also experimentally validated revealing 98.51% & 97.37% accuracies, with only 0.09% and 0.63% accuracy degradation versus the baseline models.

Conclusions

We presented and experimentally demonstrated a channel response-aware PNN designed for classifying the MNIST dataset. The PNN was trained based on the channel response of an integrated COLN[12] and its performance was experimentally validated at 20 and 25Gbaud, revealing accuracies of 98.51% & 97.37%, respectively. This work proves experimentally the ability of DL inspired training frameworks to compensate the limited channel response of photonic circuitry, paving the way for ultra-fast PNN implementations.

Acknowledgements

The work was in part funded by the EU-projects PlasmoniAC (871391) and NEBULA (871658).

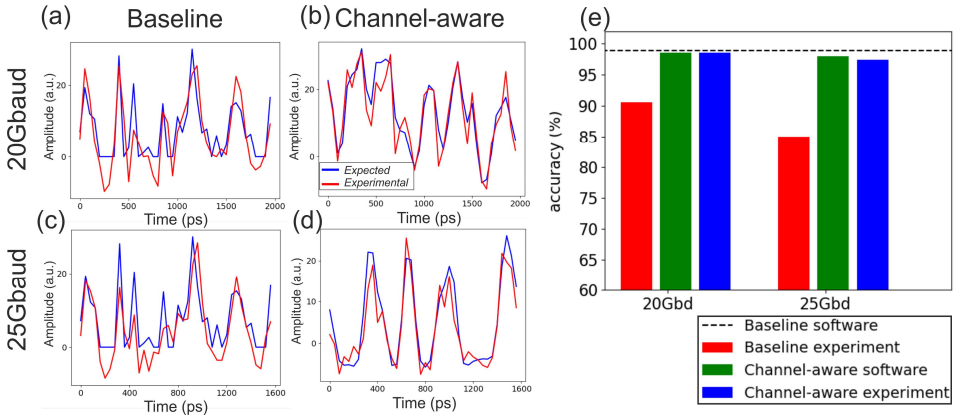

Fig. 4: Time traces of the received and NN data of the : (a),(c) Baseline model at 20 and 25 Gbaud (b),(d) Channel-aware model at 20 and 25 Gbaud (e) Accuracy results for both models at 20 and 25 Gbaud

References

[1] 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.*, vol. 26, no. 5, pp. 1–15, Sep. 2020.

[2] 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.*, vol. 26, no. 1, p. 1, 2020.

[3] 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*, 2019, vol. 1, no. c, p. SM3N.3.

[4] A. N. Tait *et al.*, “Silicon Photonic Modulator Neuron,” *Phys. Rev. Appl.*, vol. 10, no. 1, p. 1, 2019.

[5] A. N. Tait *et al.*, “Neuromorphic photonic networks using silicon photonic weight banks,” *Sci. Rep.*, vol. 7, no. 1, pp. 1–10, 2017.

[6] G. Mourgias-Alexandris, N. Passalis, G. Dabos, A. Totovic, A. Tefas, and N. Pleros, “A Photonic Recurrent Neuron for Time-Series Classification,” *J. Light. Technol.*, vol. 39, no. 5, pp. 1340–1347, 2021.

[7] 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.*, vol. 24, no. 6, pp. 1–15, 2018.

[8] Y. Shen *et al.*, “Deep learning with coherent nanophotonic circuits,” *Nat. Photonics*, vol. 11, no. 7, pp. 441–446, Jun. 2017.

[9] G. Mourgias-Alexandris *et al.*, “Neuromorphic Photonics With Coherent Linear Neurons Using Dual-IQ Modulation Cells,” *J. Light. Technol.*, vol. 38, no. 4, pp. 811–819, Feb. 2020.

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

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

[12] G. Mourgias-Alexandris, M. Moralis-Pegios, S. Simos et al., "A Silicon Photonic Coherent Neuron with 10GMAC/sec processing line-rate ", Optical Fiber Communication Conference (OFC), ID : Tu5H.1, 2021.

[13] S. Garg, A. Jain, J. Lou, and M. Nahmias, “Confounding Tradeoffs for Neural Network Quantization,” 2021.

[14] 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)*, 2019, pp. 1483–1487.

[15] 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.*, vol. 2021-January, pp. 1556–1560, 2021.

[16] N. Passalis, G. Mourgias-alexandris, A. Tsakyridis, N. Pleros, and A. Tefas, “Training Deep Photonic Convolutional Neural Networks with Sinusoidal Activations,” no. Dl, pp. 1–10.