Silicon Photonic Neuromorphic Computing with 16 GHz Input Data and Weight Update Line Rates

Apostolos Tsakyridis\(^{(1)}\), George Giamougiannis\(^{(1)}\), George Mourgas-Alexandris\(^{(1)}\), Angelina Totovic\(^{(1)}\), George Dabos\(^{(1)}\), Nikolaos Passalis\(^{(1)}\), Manos Kirtas\(^{(1)}\), Anastasios Tefas\(^{(1)}\), Miltiadis Moralis-Pegios\(^{(1)}\), Nikos Pleros\(^{(1)}\)

\(^{(1)}\)Department of Informatics, Aristotle University of Thessaloniki, 54124, Thessaloniki, Greece
ataskyridis@cad.auth.gr

Abstract: We experimentally demonstrate a silicon photonic neuron that supports on-chip input-data and weight update rates at 16GHz. Its computational performance is evaluated via the classification of the MNIST dataset achieving a mean accuracy of 99.18%.

1. Introduction

Optical Neural Networks (NN) are a prospective candidate to undertake the processing of the upsurge computational data [1] mainly due to their THz bandwidth and low latency values. Additionally, their footprint efficiency as well as their versatility to NN topologies comprise significant factors for their embodiment in future NN accelerators. As of today, the datasets favored for testing allowed for small scale NN architectures that could be mapped in a practical photonic chip [2-3], while the weight values were statically imprinted on the respective photonic weighting elements. However, the resolution of more sophisticated applications (e.g. on-chip training) would require higher dimension NNs and GHz-scale weight update rates, inducing additional burdens on their imprint to current photonic hardware.

In this work, we present the feasibility of executing two cascaded NN layers over a smaller dimension elementary photonic neural layer via the employment of Time Division Multiplexing (TDM) and fast weight update. Harnessing the high bandwidth offered by the SiGe Electro Absorption Modulators (EAMs) [4], we TDM both the input and weight values of a neural layer and imprint them on-chip via the respective EAMs driven at the rate of 16 GHz, projecting the versatility of a compact elementary block to implement high dimension NN tasks. The computational efficiency of the photonic hardware was experimentally benchmarked via the classification of the MNIST dataset achieving an accuracy of 99.18%. To the best of our knowledge, it is the first time that input data and weights can be simultaneously updated on a photonic neuron at this record-high rate, paving the way for on-chip photonic computing to be utilized in a set of vast new applications in machine learning tasks, i.e., NN training, real-time classification.

2. Architecture, experimental setup and results

The deployed 2-input SiPho neuromorphic prototype, that employs the coherent neuron architecture analyzed in [4-5], is depicted in Fig. 1(a). In order to assess its computational efficiency, we designed a NN-based classifier, with the final two NN layers deployed over the SiPho hardware. More specifically, the 5-layer NN depicted in Fig. 1(b) was trained via software for the recognition of the handwritten digits 3 and 5 of the MNIST dataset. The last two layers, highlighted in the light green box of Fig. 1(b), that were designed to be composed of 2-input neurons, were implemented in the photonic domain via our coherent linear neuron. More precisely, the four 1902-length inputs of

![Fig. 1](image-url)
the penultimate layer \(x_1-x_4\), along with their respective weights \(w_1-w_4\) were multiplexed to form the corresponding \(x_{a_1}w_{a_1}\) and \(x_{b_1}w_{b_1}\) sequences that were utilized to drive the respective EAMs. Specifically, we multiplexed, in the time domain, the input data of the 4 neurons into two sequences, with \(x_a\) and \(x_b\) being equal to the streams \(X_1,X_2,X_3,...X_1902X_1902\) and \(X_2,X_4,X_2,X_4,...X_1902X_1902\), respectively. The same multiplexing scheme was applied to their respective weights with \(w_a\) and \(w_b\) being equal to \(w_1w_1w_1w_1\) and \(w_2w_2w_2w_2\) respectively. A visualization of the above described TDM scheme is illustrated in Fig. 1(c). The fast imprinting rate of the input and weight values to the photonic hardware, that our prototype supports, enables the employment of the TDM technique, allowing in this way the one-step implementation of a whole 4:2 NN feed-forward layer via the elementary 2-input neuron block. The experimental setup used is depicted in Fig. 1(d). A light beam at 1560 nm was injected to the SiPho chip via a pdk-ready TE grating coupler. An arbitrary waveform generator was utilized to generate the \(x_{a_1}\), \(w_{a_1}\) and \(x_{b_1}\), \(w_{b_1}\) sequences, at 16 GHz that fed the input and weight EAMs of the upper and the lower branch of the MZI. The multiplexed weighted summations were coupled out of the chip, injected to a photodiode, captured by a real time oscilloscope and demultiplexed to the \(\Sigma_1\) and \(\Sigma_2\) weighted summations. The softwarely activated linear summations were then transferred to the photonic domain in order to realize the output layer of the NN. Fig. 2 illustrates the results obtained during the experimental evaluation of our prototype for the classification of the MNIST dataset. More specifically, Fig. 2(a) and (c) depict the experimental (blue lines) time traces of the time division multiplexed streams \(x_{a_1}\), \(x_{b_1}\) at 16 GHz, along with their respective reference (orange lines). Their mean squared error (MSE) was measured to be 0.033 and 0.082, respectively. The traces of their corresponding weights \(w_{a_1}\), \(w_{b_1}\) are shown in Fig. 2(b) and (d). It can be observed in Fig. 2(b) that the weight values of \(w_1\) alternate with the weight values of \(w_3\) every 62.5 ps (1/16 GHz), formulating the multiplexed sequence \(w_{1}\) with an MSE of 0.003. The same performance was measured for \(w_2\). Fig. 2(e) and (f) illustrate the time traces of the demultiplexed weighted summations performed in the two neurons of the penultimate layer that were obtained with an MSE of 0.001 and 0.006 in the upper \(\Sigma_1\) and lower \(\Sigma_2\) neuron, respectively. Finally, Fig. 2(g) depicts the time trace of the output layer that yields an MSE of 0.019. The accuracy distribution of the classification of the MNIST dataset is depicted in the histogram of Fig. 2(h) with its mean value being equal to 99.18%, while the targeted accuracy that was achieved via the software was 99.47%.

3. Conclusion

We presented and experimentally demonstrated an optical coherent neuron classifying the MNIST dataset with a mean accuracy of 99.18%, enabling both input data and weight imprinting at 16 GHz. Additionally, by employing TDM we tested a whole neural layer in a single step, via an elementary neuron block integrated in a compact SiPho chip, projecting its versatility to implement any-dimensional neural layer or network.

Acknowledgments

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

References