PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

Photonic tensor core for machine learning: a review

Nicola Peserico, Xiaoxuan Ma, Bahvin Shastri, Volker Sorger

Nicola Peserico, Xiaoxuan Ma, Bahvin J. Shastri, Volker J. Sorger, "Photonic tensor core for machine learning: a review," Proc. SPIE 12204, Emerging Topics in Artificial Intelligence (ETAI) 2022, 1220407 (3 October 2022); doi: 10.1117/12.2633916

SPIE.

Event: SPIE Nanoscience + Engineering, 2022, San Diego, California, United States

Invited Paper

Photonic Tensor Core for Machine Learning: a review

Nicola Peserico[a], Xiaoxuan Ma[a], Bhavin J. Shastri[b], and Volker J. Sorger[a]

[a]The George Washington University, 800 22nd st NE, Washington DC, USA
[b]Queens University, Kingston, Ontario, Canada

ABSTRACT

Photonic Tensor Core circuits have been widely explored as possible hardware accelerators for the next generation of Machine Learning applications, due to the large bandwidth, low latency, and energy saving that light has. Many architectures have been presented, especially exploiting photonic integrated circuits. However, most of the proposed solutions lack some features, such as integration, scalability, or energy saving. In this paper, we review the major achievements in recent years, showing how high integration can lead to better performance, but it could also limit the scalability of the overall system.

Keywords: Photonic Tensor Core, Machine Learning, Neural Network, Silicon Photonics

1. INTRODUCTION

Artificial Intelligent has raised as the technology of this century. Its main actor, Machine Learning (ML), has become one of the major technologies of the last decade, thanks to the exponential growth of the Deep Learning and Neural Networks (NN), that had achieved major results from gaming,[1] self-driving,[2] up to protein prediction in the biology field.[3] This massive boost in the ML applications has been possible by the improvements in the hardware field, where processors, such as GPUs,[4] are capable to perform millions of products and sums in a much shorter amount of time compared to decades ago, reducing the time needed for training and execution of the Deep Learning algorithms.

Neural Networks in particular require a great effort on the computational side due to the specific structure of the neuron: the output of a single neuron in a layer is formed from the output of all the previous layer neurons, scaled by certain trainable weights, and passed through an activation function. As we can see, the structure can be decomposed into 2 main parts, a linear vector-vector multiplication (or matrix-vector considering an entire layer), and a non-linear part formed by the activation function. Deep Learning leverages this NN scheme by building pipelines with several layers, whose size can be in the orders of hundreds. Processors that are specifically designed to excel in those tasks have a clear edge to respect general processors such as common CPUs, and this has been leveraged up to the design of application-specific integrated circuits (called ASICs) to perform just NN tasks.

However, those processors still rely on digital electronics to perform the tasks, and that results in many disadvantages, such as high latency, low speed, and high power consumption.[5] Analog electronic solutions have been proposed in the latest years to address these limitations, with FPGA solutions,[6] PCM,[7] memsistor,[8] and protonic approaches.[9] But the underlying limits of the electron to work at high speed in an energy-efficient way still stay.

Optics and photonics, on the other hand, can overcome those limitations by relying on the electromagnetic behavior of the light to perform neural network tasks. Exploiting the interference of the light, it is possible to perform Multiplication and Accumulation (MAC) operations, as well as nonlinear functions using proper circuits or materials. These operations can be performed in an almost energy-free fashion, as just a small portion of the light is lost by adsorption or scattering.[10] Moreover, Photonic Integrated Circuits (PICs) support large bandwidth, up to 100GHz, working with the lowest latency achievable by photon propagation.[11]

In this paper, we review the main architectures, devices, and achievements in implementing the tasks required by any Neural Network on a PIC. First, we highlight the main 2 architectures used, based on coherent single-source

Further author information: (Send correspondence to V.S)
A.A.A.: E-mail: sorger@gwu.edu

Emerging Topics in Artificial Intelligence (ETAI) 2022, edited by Giovanni Volpe, Joana B. Pereira, Daniel Brunner, Aydogan Ozcan, Proc. of SPIE Vol. 12204, 1220407 · © 2022 SPIE · 0277-786X · doi: 10.1117/12.2633916

interference and multiple-wavelength incoherent one. We then focus on the integration of the activation function, showing the pros and cons of different solutions. The conclusion will focus on the level of integration achieved and the next step that this field can bring to the scientific community.

2. ARCHITECTURES

Photonic Tensor Core can be implemented in multiple ways and architectures,[12] here we look at the two most major architectures for Silicon Photonics, based on the mathematical approach used.

2.1 Coherent PTC

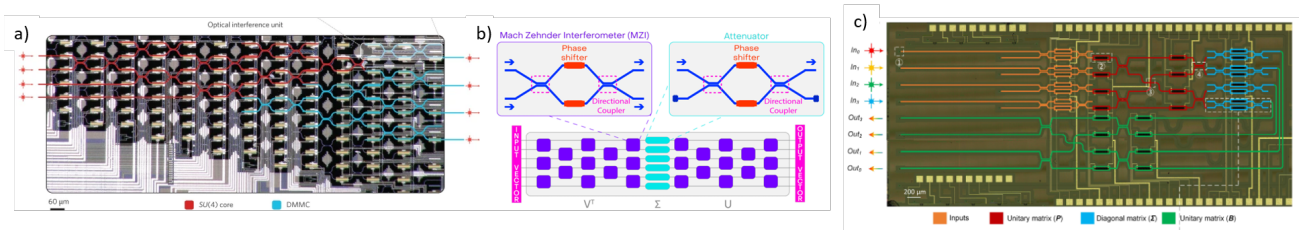

Figure 1. Here presented some examples of SVD-based Silicon Photonics PTC. (a) The first solution presented by Shen et al,[13] consists of a mesh of MZI whose task is to compute the matrix multiplication. (b) The architecture reflects the one presented by Lightmatter, where the decomposition is visible.[14] (c) The last one shows the possibility to implement the same SVD approach by using multiple laser sources, as presented by Fend et al.[15]

The first group we review includes all the PTC implementations that use coherent interference to perform the MAC operation and Matrix multiplication in a wider sense. The major used scheme relies on the Single Value Decomposition (SVD), a mathematical property of the matrix to be decomposed into two unitary matrices and a diagonal one. This property allows for the implementation of matrix multiplication by using a mesh of Mach-Zehnder Interferometers (MZIs) and one single laser source. The framework for this implementation is based on the work of Miller et al.[16,17]

Initial experimental implementations were done by different groups, focusing on different applications. Annoni et al. implemented an MZI network for mode un-scrambling,[18] while the first implementation for Deep Learning was realized by Shen et al.[13] This last implementation shows the possibility to implement a multi-layer optical NN for vowel recognition, showing an initial 76.7% experimental accuracy, that could be further boosted by fine-tuning of the MZI.

As shown in figure 1, recent implementations have been published by both universities and industries, following a push for accelerators for NN. One major step forward has been achieved by Lightmatter, showing the steps forward in the integration of the PIC with the electronic system needed to control and transfer the data.[14,19] Another examples of the same architecture was presented by Feng et al, where the PIC was integrated into a butterfly electronic board, showing a good accuracy with just 3-bit weights precision,[15] and by Zhang et al, that have implemented a complex-value version for NN.[20] Last example of a coherent system has been proposed by Giamougiannis et al., where a novel interferometric coherent photonic crossbar architecture steps forward the common SVD, but still maintains the single laser source.[14] Same circuit has been used by Lau et al.,[21] to predict the quantum mechanical properties of molecules. The circuit has shown a its robustness also in more complex algoriths, for example Zhou et al. have used it to run PageRank algorithm.[22] A similar approach has been proposed by Youngblood et al.,[23] for a full matrix-matrix multiplication scheme. In all these solutions, however, the activation function is performed by a digital computational unit, separate from the optical circuit.

Moreover, controlling the MZI requires controlling the phase matching of both incoming inputs, as well as the phase difference between the 2 arms. In recent work, Banerjee et al. have shown the impact o the uncertainties in this type of NN, and how those uncertainties can affect the final results, and so the accuracy of the system.[24]

2.2 Signal Multiplexing PTC

Another approach to perform the matrix multiplication on a PTC is by directly pairing the matrix weights with matrix tunable elements on the chip, exploiting some of the possible multiplexing schemes that photonics

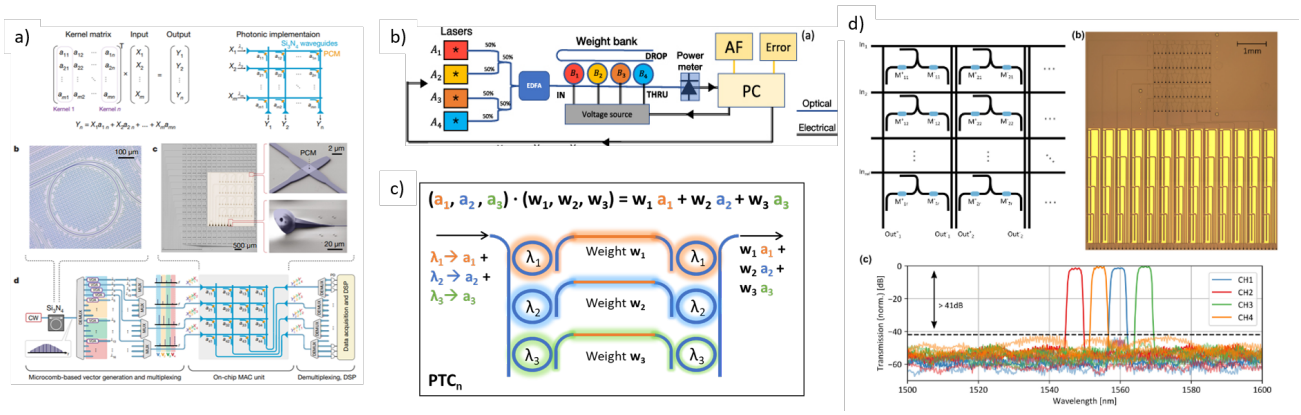

Figure 2. Here are shown some of the architectures that rely on multiple laser sources to perform MAC operations. (a) The first scheme presented is based on a WDM grid using tunable cross-bar couplers as weights.[25] (b) Another approach has been proposed by using microring resonators as WDM weights, by Marquez et al.[26] (c) Similar to the previous, Ma et al. have presented the possibility to separate WDM (de-)multiplexing from the weighting, allowing to select the proper weight components for the application[27] (d-e) Last one example is presented by Bruckerhoff et al. exploiting a broadband architecture using Phase Change Material.[28]

has, such as time, wavelength, or mode multiplexing. There are several solutions to implement this approach, varying the components that are performing the weights. In most cases, the input vector is formed by a set of amplitude modulated wavelengths, that are injected into the chip separately or combined. The task of the PIC is to combine the input vector in a different configuration, depending on the weight that each input has on the specific output port.

The first architecture has been proposed by Feldmann et al.,[25] and it exploits the cross-bar coupling to mix the incoming separate input wavelengths into one single channel. This solution uses Phase Change Materials (PCMs) to adjust the weights, allowing for high energy savings, and high throughput, up to 10^{12} MAC operations per second. The choice of using PCMs has the drawback of limiting the updating of the weights, by so limiting the possibility to perform training on-chip. Similar to this scheme, Bruckerhoff et al. have proposed a cross-bar solution with ultra-low crosstalk and high bandwidth, using GST as PCM material, and showing the crosstalk to be -40dB than the actual signal.[28]

Another approach is by using microring resonators, as or mux/demux, either directly as weights.[29] The first implementation for NN by Tait et al.[30] has shown the possibility to use this scheme as PTC, with an accuracy in the weights of 5.1 bit at 2Gbps signals. Marquez et al. have implemented by demonstrating a Hopfield network on chip, an example of a recurrent neural network, obtaining high accuracy in pattern recognition.[26] The same scheme has been leveraged to achieve an even higher weights accuracy of 9 bit, by using low-speed labeling.[31]

An alternative scheme has been proposed by Miscuglio et al.,[32] where the weights are implemented between a series of demux and mux add-drop microring resonators. This architecture has been proven by Ma et al.[27] with tunable low-speed MZI acting as weights, and by using PCM components.[33] This double approach permits to adap of t the circuit to the applications of the NN, for example, tunable high-speed weights can be used in Cloud applications, where training requires a high rate of updates, while the PCM solutions, exploiting the non-volatility, can be adapt to edge computing, where energy efficiency is a major concern.[34] Recently, a similar scheme has been used by Sarwat et al. to demonstrate unsupervised correlation detection, using GST as PCM material acting as weights.[35]

By using WDM, most of these schemes require more electronics to monitor and control the various tunable elements, such as weights, microrings, and so on. These feedback control can be performed by heaters and proper PID circuits, but it would impact the energy efficiency of the PTC, limiting its application spectrum.

3. ACTIVATION FUNCTION

Up to now, all the schemes we have seen do not implement the activation function on the chip, since optics do not have a straightforward way to implement nonlinear behaviors. Recent works have proposed a different way to perform such a function, exploiting either an O/E/O pipeline or integrated electro-optical circuits. The

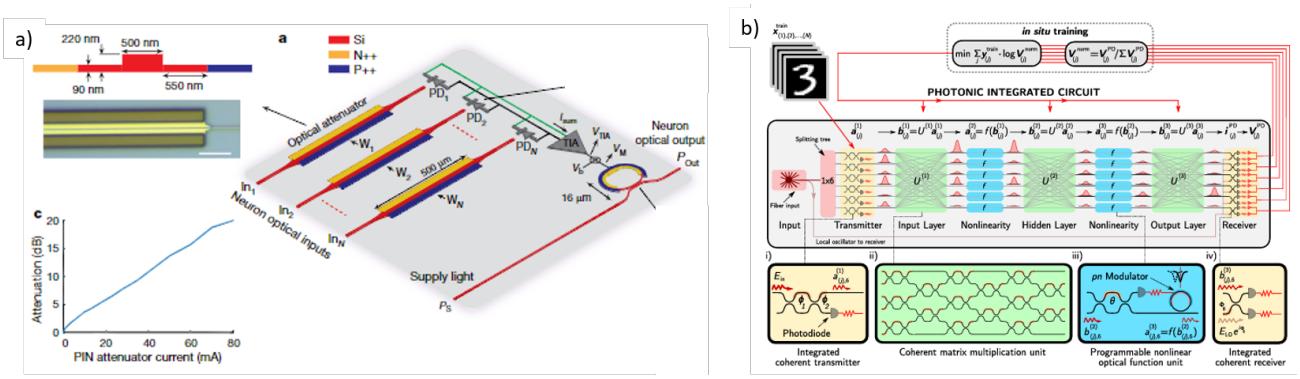

Figure 3. Here are shown two major schemes to implement the activation function on-chip. (a) The first one, proposed by Ashtiani et al,[36] comprises the complete domain change of all the incoming signals from the optical to electrical domain and then using this signal to pilot a microring resonator for the next neuron layer. (b) In the alternative scheme, proposed by Bandyopadhyay et al,[37] uses just part of the signal to pilot a modulator that is acting directly on the original signal.

first implementation is based on the complete domain transition of the optical signal to the electrical one, to pilot the following optical element. Huang et al.[38] have implemented this method to construct a three-layer fully-connected NN to compensate for the non-linearity of the light propagation in the fiber. A similar approach has been made by Ashtiani et al.,[36] where the actual accumulation is performed electrically, by summing the currents generated by the different photodetectors, one per each incoming signal. In this case, the authors presented a fully integrated 3-layers NN, demonstrating high accuracy (up to 93.8%) and impressive low latency, down to 570 ps. Another strategy was proposed by Feldmann et al.,[39] where PCM materials are activated by optical spiking generated by the WDM sum of incoming input signals. The circuit is capable of supervised and unsupervised learning, showing one of the first photonic neurosynaptic networks.
Another solution has been implemented by Bandyopadhyay et al,[37] in a coherent PTC scheme. In this case, the activation function is piloted by a small fraction of the incoming signal that modulates the resonance position of a microring resonator, feed by the rest of the signal itself. In this scheme, the same signal out of one layer is directly sent into the following neural layer, without the need for additional sources or modulators. Be noticed, that this scheme can't be straightforwardly implemented in a WDM fashion, since the WDM weights would disassemble the signal coming from the previous layer, losing its information. In this work, a full 3-layer NN is implemented in a $6x5.7mm^2$ PIC, having 169 active devices on the same chip. It has been shown the NN can perform vowel classification with 92.7% accuracy.

In all the solutions proposed, the main limit is the fixed size of the NN, in terms of the number of layers and number of neurons per layer. Since the non-linear behavior of the activation function, strategies like the GeMM compiler, where the matrix multiplication is performed by splitting the matrix into smaller ones,[40] is not possible anymore, reducing the scalability of the overall system. Moreover, since the current generated by the photodetector is low, it can not directly modulate an optical signal, requiring to add amplification stage to reach the proper VπL. To improve this aspect, and by so reducing, even more, energy consumption, one possibility is to switch to ITO-based modulators, that can achieve VπL as low as 95Vμm,[41–44] or ITO-graphene devices, that can work with a bandwidth of over 130GHz.[45] The strategy of mixing multiple materials to achieve on-chip activation function was discussed by Miscuglio et al.,[46] where a road-map identify the major challanges were proposed.

4. CONCLUSION

In this paper, we review the major and recent architectures to implement an optical Neural Network, dividing the initial linear part regarding the Photonic Tensor Core, as an accelerator to the MAC operations, and a second part dedicated to the strategies used to integrate the activation function on-chip. For the linear PTC part, 2 main approaches are now being investigated, both at the academic and commercial levels. The main parameters and Figure-of-Merit that set the challenges are the MAC operations per second, the MAC operation per Energy, the total number of active devices, and the total size of the PIC. Some trade-offs can be seen, as using P-RAM PCM elements can improve the energy consumption,[47] but can limit the updating rates. On the other hand, the implementation of the activation function has shown some interesting progress in recent times, pointing to several possibilities to implement such a function in the PTC itself. However, still, major achievements must be addressed, like the fixed size of the Neural Network, and the energy requirement for piloting the integrated p-i-n junction modulators. In the future, we will see the first commercial applications of such technologies, as well as more diffusion in the data centers, as the request for Machine Learning applications and accelerators will continue to grow.

ACKNOWLEDGMENTS

V.S.J. is supported by the PECASE under AFOSR contract (FA9550-20-1-0193)

REFERENCES

[1] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Driessche, G. V. D., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D., "Mastering the game of go with deep neural networks and tree search," *Nature* **529** (2016).

[2] Grigorescu, S., Trasnea, B., Cocias, T., and Macesanu, G., "A survey of deep learning techniques for autonomous driving ai for self-driving vehicles, artificial intelligence, autonomous driving, deep learning for autonomous driving," *J Field Robotics* **37**, 362–386 (2020).

[3] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., Hassabis, D., and Hassabis, D., "Highly accurate protein structure prediction with alphafold," *Nature* **596**, 583 (2021).

[4] Strigl, D., Kofler, K., and Podlipnig, S., "Performance and scalability of gpu-based convolutional neural networks," *2010 18th Euromicro Conference on Parallel, Distributed and Network-based Processing* (2010).

[5] Merolla, P. A., Arthur, J. V., Alvarez-Icaza, R., Cassidy, A. S., Sawada, J., Akopyan, F., Jackson, B. L., Imam, N., Guo, C., Nakamura, Y., Brezzo, B., Vo, I., Esser, S. K., Appuswamy, R., Taba, B., Amir, A., Flickner, M. D., Risk, W. P., Manohar, R., and Modha, D. S., "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science* **345**, 668–673 (8 2014).

[6] Guo, K., Zeng, S., Yu, J., Wang, Y., and Yang, H., "A survey of fpga-based neural network accelerator," (12 2017).

[7] Khaddam-Aljameh, R., Member, G. S., Stanisavljevic, M., Mas, J. F., Karunaratne, G., Brändli, M., Liu, F., Singh, A., Müller, S. M., Member, S., Egger, U., Petropoulos, A., Antonakopoulos, T., Brew, K., Choi, S., Ok, I., Lie, F. L., Saulnier, N., Chan, V., Ahsan, I., Narayanan, V., Nandakumar, S. R., Gallo, M. L., Francese, P. A., Sebastian, A., Eleftheriou, E., Fellow, L., and are, A. S., "Hermes-core-a 1.59-tops/mm 2 pcm on 14-nm cmos in-memory compute core using 300-ps/lsb linearized cco-based adcs," *IEEE JOURNAL OF SOLID-STATE CIRCUITS* **1**.

[8] Ankit, A., Hajj, I. E., Chalamalasetti, S. R., Ndu, G., Foltin, M., Williams, R. S., Faraboschi, P., Hwu, W.-m. W., Strachan, J. P., Roy, K., et al., "Puma: A programmable ultra-efficient memristor-based accelerator for machine learning inference," in [*Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*], 715–731 (2019).

[9] Onen, M., Emond, N., Wang, B., Zhang, D., Ross, F. M., Li, J., Yildiz, B., and del Alamo, J. A., "Nanosecond protonic programmable resistors for analog deep learning," *Science* **377**, 539–543 (7 2022).

[10] Miller, D. A. B., "Device requirements for optical interconnects to silicon chips; device requirements for optical interconnects to silicon chips," *Proceedings of the IEEE* **97** (2009).

[11] Sun, S., Narayana, V. K., Miscuglio, M., Kimerling, L. C., El-Ghazawi, T., Volker, ., and Sorger, J., "clear: A holistic figure-of-merit for post-and predicting electronic and photonic-based compute-system evolution,"

[12] Zhou, H., Dong, J., Cheng, J., Dong, W., Huang, C., Shen, Y., Zhang, Q., Gu, M., Qian, C., Chen, H., Ruan, Z., and Zhang, X., "Photonic matrix multiplication lights up photonic accelerator and beyond," *Light: Science & Applications 2022 11:1* **11**, 1–21 (2 2022).

[13] Shen, Y., Harris, N. C., Skirlo, S., Prabhu, M., Baehr-Jones, T., Hochberg, M., Sun, X., Zhao, S., Larochelle, H., Englund, D., and Soljačić, M., "Deep learning with coherent nanophotonic circuits," *NATURE PHOTONICS* — **11** (2017).

[14] Demirkiran, C., Eris, F., Wang, G., Elmhurst, J., Moore, N., Harris, N. C., Basumallik, A., Reddi, V. J., Joshi, A., and Bunandar, D., "An electro-photonic system for accelerating deep neural networks," (9 2021).

[15] Feng, C., Gu, J., Zhu, H., Ying, Z., Zhao, Z., Pan, D. Z., and Chen, R. T., "A compact butterfly-style silicon photonic-electronic neural chip for hardware-efficient deep learning," (2021).

[16] Miller, D. A. B., "Self-configuring universal linear optical component [invited]," *Photonics Research, Vol. 1, Issue 1, pp. 1-15* **1**, 1–15 (6 2013).

[17] Miller, D. A. B., "Perfect optics with imperfect components," *Optica, Vol. 2, Issue 8, pp. 747-750* **2**, 747–750 (8 2015).

[18] Annoni, A., Guglielmi, E., Carminati, M., Ferrari, G., Sampietro, M., Miller, D. A., Melloni, A., and Morichetti, F., "Unscrambling light—automatically undoing strong mixing between modes," *Light: Science & Applications 2017 6:12* **6**, e17110–e17110 (6 2017).

[19] Ramey, C., "Silicon photonics for artificial intelligence acceleration: Hotchips 32," in [*2020 IEEE hot chips 32 symposium (HCS)*], 1–26, IEEE (2020).

[20] Zhang, H., Gu, M., Jiang, X. D., Thompson, J., Cai, H., Paesani, S., Santagati, R., Laing, A., Zhang, Y., Yung, M. H., Shi, Y. Z., Muhammad, F. K., Lo, G. Q., Luo, X. S., Dong, B., Kwong, D. L., Kwek, L. C., and Liu, A. Q., "An optical neural chip for implementing complex-valued neural network,"

[21] Lau, J. W. Z., Zhang, H., Wan, L., Shi, L., Cai, H., Luo, X., Lo, P., Lee, C.-K., Kwek, L.-C., and Liu, A. Q., "A photonic chip-based machine learning approach for the prediction of molecular properties," *arXiv preprint arXiv:2203.02285* (2022).

[22] Zhou, H., Zhao, Y., Xu, G., Wang, X., Tan, Z., Dong, J., and Zhang, X., "Chip-scale optical matrix computation for pagerank algorithm; chip-scale optical matrix computation for pagerank algorithm," *IEEE Journal of Selected Topics in Quantum Electronics* **26**, 8300910 (2020).

[23] Youngblood, N., "Coherent photonic crossbar arrays for large-scale matrix-matrix multiplication; coherent photonic crossbar arrays for large-scale matrix-matrix multiplication," *IEEE Journal of Selected Topics in Quantum Electronics* **PP** (2022).

[24] Banerjee, S., Nikdast, M., and Chakrabarty, K., "On the impact of uncertainties in silicon-photonic neural networks; on the impact of uncertainties in silicon-photonic neural networks," *IEEE Design & Test* **PP**, 2168–2356 (2022).

[25] Feldmann, J., Youngblood, N., Karpov, M., Gehring, H., Li, X., Stappers, M., Gallo, M. L., Fu, X., Lukashchuk, A., Raja, A. S., Liu, J., Wright, C. D., Sebastian, A., Kippenberg, T. J., Pernice, W. H. P., and Bhaskaran, H., "Parallel convolutional processing using an integrated photonic tensor core," *Nature* **589** (2021).

[26] Liu, G., Ma, W.-P., Cao, H., al, Zhu, R., Qiu, T., Wang, J., Marquez, B. A., Guo, Z., Morison, H., Shekhar, S., Chrostowski, L., Prucnal, P., and Shastri, B. J., "Photonic pattern reconstruction enabled by on-chip online learning and inference," *Journal of Physics: Photonics* **3**, 024006 (2 2021).

[27] Ma, X., Peserico, N., Khaled, A., Guo, Z., Nouri, B., Llc, O., Dalir, H., Shastri, B., Sorger, V., Washington, G., Guo, Z., Nouri, B. M., Shastri, B. J., and Sorger, V. J., "High-density integrated photonic tensor processing unit with a matrix multiply compiler," (7 2022).

[28] Brückerhoff-Plückelmann, F., Feldmann, J., Gehring, H., Zhou, W., Wright, C. D., Bhaskaran, H., and Pernice, W., "Broadband photonic tensor core with integrated ultra-low crosstalk wavelength multiplexers," *Nanophotonics* **11**, 4063–4072 (9 2022).

[29] Ding, J., Zhang, L., Yang, L., Xu, Q., and Ji, R., "On-chip cmos-compatible optical signal processor," *Optics Express, Vol. 20, Issue 12, pp. 13560-13565* **20**, 13560–13565 (6 2012).

[30] Tait, A. N., Jayatilleka, H., Ferreira, T., Lima, D. E., Ma, P. Y., Nahmias, M. A., Shastri, B. J., Shekhar, S., Chrostowski, L., and Prucnal, P. R., "Feedback control for microring weight banks," *Optics Express, Vol. 26, Issue 20, pp. 26422-26443* **26**, 26422–26443 (10 2018).

[31] Zhang, W., Huang, C., Huang, C., Peng, H.-T., Bilodeau, S., Jha, A., Blow, E., de Lima, T. F., de Lima, T. F., Shastri, B. J., Shastri, B. J., and Prucnal, P., "Silicon microring synapses enable photonic deep learning beyond 9-bit precision," *Optica, Vol. 9, Issue 5, pp. 579-584* **9**, 579–584 (5 2022).

[32] Miscuglio, M. and Sorger, V. J., "Photonic tensor cores for machine learning," *Applied Physics Reviews* **7**, 031404 (2 2020).

[33] Ma, X., Meng, J., Peserico, N., Miscuglio, M., Zhang, Y., Hu, J., and Sorger, V. J., "Photonic tensor core with photonic compute-in-memory," *Optical Fiber Communication Conference (OFC) 2022 (2022), paper M2E.4* , M2E.4 (3 2022).

[34] Peserico, N., de Lima, T. F., de Lima, T. F., Prucnal, P., and Sorger, V. J., "Emerging devices and packaging strategies for electronic-photonic ai accelerators: opinion," *Optical Materials Express, Vol. 12, Issue 4, pp. 1347-1351* **12**, 1347–1351 (4 2022).

[35] Sarwat, S. G., Brückerhoff-Plückelmann, F., Carrillo, S. G. C., Gemo, E., Feldmann, J., Bhaskaran, H., Wright, C. D., Pernice, W. H., and Sebastian, A., "An integrated photonics engine for unsupervised correlation detection," *Science Advances* **8**, 3243 (6 2022).

[36] Ashtiani, F., Geers, A. J., and Aflatouni, F., "An on-chip photonic deep neural network for image classification," *Nature* **606** (2022).

[37] Bandyopadhyay, S., Sludds, A., Krastanov, S., Hamerly, R., Harris, N., Bunandar, D., Streshinsky, M., Hochberg, M., and Englund, D., "Single chip photonic deep neural network with accelerated training," (2022).

[38] Huang, C., Fujisawa, S., de Lima, T. F., Tait, A. N., Blow, E. C., Tian, Y., Bilodeau, S., Jha, A., atih Yaman, F., Peng, H.-T., Batshon, H. G., Shastri, B. J., Inada, Y., Wang, T., and Prucnal, P. R., "Silicon photonic-electronic neural network for fibre nonlinearity compensation," *Nature Electronics* **4**, 837–844 (10 2021).

[39] Feldmann, J., Youngblood, N., Wright, D., Bhaskaran, H., and Pernice, W. H. P., "All-optical spiking neurosynaptic networks with self-learning capabilities photonic implementation of an artificial neuron," *Nature* .

[40] Guo, Z., Tait, A. N., Marquez, B. A., Filipovich, M., Morison, H., Prucnal, P. R., Fellow, L., Chrostowski, L., Member, S., Shekhar, S., Shastri, B. J., Guo, Z. G. Z., and Mori, H., "Multi-level encoding and decoding in a scalable photonic tensor processor with a photonic general matrix multiply (gemm) compiler," *IEEE JOURNAL OF SELECTED TOPICS IN QUANTUM ELECTRONICS* **28**, 8300714.

[41] Amin, R., George, J. K., Sun, S., Lima, T. F. D., Tait, A. N., Khurgin, J. B., Miscuglio, M., Shastri, B. J., Prucnal, P. R., El-Ghazawi, T., and Sorger, V. J., "Ito-based electro-absorption modulator for photonic neural activation function articles you may be interested in," **7**, 81112 (2019).

[42] Amin, R., Maiti, R., Gui, Y., Suer, C., Miscuglio, M., Heidari, E., Khurgin, J. B., Chen, R. T., Dalir, H., Volker, ., and Sorger, J., "Heterogeneously integrated ito plasmonic mach-zehnder interferometric modulator on soi," *Scientific Reports* — **11**, 1287 (123).

[43] Gui, Y., Nouri, B. M., Miscuglio, M., Amin, R., Wang, H., Khurgin, J. B., Dalir, H., and Sorger, V. J., "100 ghz micrometer-compact broadband monolithic ito mach-zehnder interferometer modulator enabling 3500 times higher packing density," *Nanophotonics* **11**, 4001–4009 (9 2022).

[44] Amin, R., Maiti, R., George, J. K., Ma, X., Ma, Z., Dalir, H., Miscuglio, M., and Sorger, V. J., "A lateral mos-capacitor-enabled ito mach–zehnder modulator for beam steering," *Journal of Lightwave Technology* **38**(2), 282–290 (2020).

[45] Amin, R., George, J. K., and Wang, H., "An ito-graphene heterojunction integrated absorption modulator on si-photonics for neuromorphic nonlinear activation collections articles you may be interested in," *APL Photonics* **6**, 120801 (2021).

[46] Miscuglio, M., Adam, G. C., Kuzum, D., and Sorger, V. J., "Roadmap on material-function mapping for photonic-electronic hybrid neural networks articles you may be interested in," **7**, 100903 (2019).

[47] Meng, J., Peserico, N., Ma, X., Zhang, Y., Popescu, C.-C., Kang, M., Miscuglio, M., Richardson, K., Hu, J., and Sorger, V., "Electrical programmable low-loss high cyclable nonvolatile photonic random-access memory," (2022).