Photonic Neuromorphic Processors and Convolutional Neural Network Accelerator

J.K. George[1], A. Mehrabian[1], R. Amin[1], T. Ferreira de Lima[2], A. N. Tait[2], B. J. Shastri[2], T. El-Ghazawi[1], P. R. Prucnal[2], V. J. Sorger[1]

[1]Department of Electrical and Computer Engineering, George Washington University, Washington DC, 20052, USA
2Department of Electrical Engineering, Princeton University, Princeton NY, 08544, USA

Summary. Here we discuss the value propositions and recent advances of photonic neuromorphic networks (NN) in the context of non-van Neumann processors to include feed-forward and convolutional NNs (CNN). We show that optics allows synergistic paradigms, namely a) natural convolution for CNNs, and b) its built-in parallelism such as in wave-length-division-multiplexing enables a broadcast-and-weight protocols for all-to-all connected recurrent NNs with record-high MAC/J efficiency and short-runtime delay (10's ps) just limited by the photon's time-of-flight.

Photonic neural networks (PNN) are promising alternatives to electronic GPUs to perform machine-learning tasks. The PNNs value proposition originates from **i)** near-zero energy consumption for vector matrix multiplication (VMM) once trained, **ii)** 10-100 ps short interconnect delays given by the photon's time-of-flight through the PNN (=runtime), and **iii)** only requiring 'weak' optical nonlinearity which can be provided via ~fJ/bit efficient emerging electrooptic devices (**Fig. 1**). Why now? Photonic integrated circuits (PIC) offer high data bandwidth at low latency, with competitive footprints and synergies to microelectronics architectures such as foundry access. Here the MAC/J depends on the E/bit of the nonlinear activations provided device-level efficiency (e.g. EOM, all-optical thresholder) reach 10^18 MAC/J using nanophotonic devices (**Fig. 1**). The PNN's speed depends on the fan-in (N_{Fi} = #synaptic connections), i.e. about 100 in WDM PNNs based-on photonic integrated circuits (PIC) [1]. Regarding PNN applications, for speech recognition (e.g. Siri, Amazon Alexa) electronic NN's are sufficiency fast. However, for nonlinear optimization, predictive control (e.g. UAV flight control), and real-time processing (e.g. warfighter data processing), photonic solutions may become relevant due to the short delay and energy efficiency (i.e. processing-at-the-edge) in an IoT edge-computing future.

NNs require both a weighting of inputs and a nonlinear activation function operating on their sum, e.g. perceptron model (**top, Fig.**

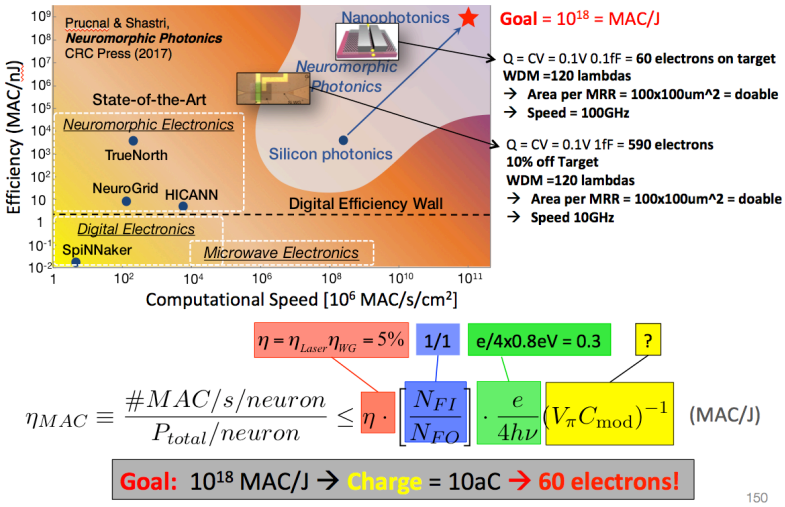

Figure 1. Performance vectors for compute systems. Photonic neuromorphic processors break the digital efficiency wall set by CMOS and von-Neumann architectures [1]. 10^18 MAC/J can be reached with photonic neural networks (PNN) with attojoule-efficient

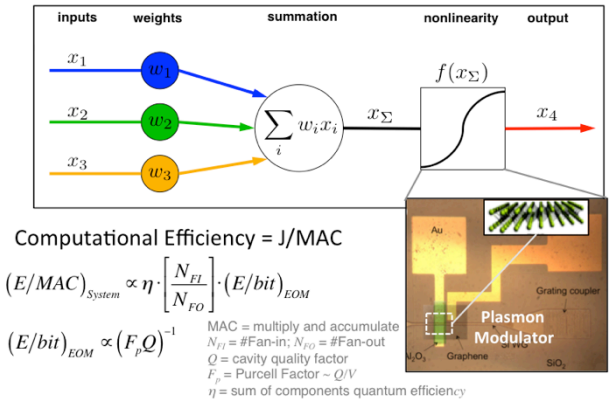

Figure 2. Model of a photonicperceptron neuron inside a PNN. The nonlinear activation function is provided by the transfer function of the EOM. The computational system efficiency scales with the modulator efficiency, which in turn scales inversely with device the physical volume, *V*, and material broadening [1,2]. Our plasmon modulators are micrometer-compact consuming 100's aJ [4,5].

2). Weighting has been demonstrated in integrated photonics with both interferometric and ring-based wavelength division multiplexing [1]. While direct nonlinearity in optics being difficult to achieve without high optical powers, electro-optic nonlinearity can be created by directly coupling a photodiode to an EOM [2-5], when coupled to the summation performing photodiode (ideally capacitively for short RC-delay [6]). Here we consider a plurality of EOMs inside such a photonic neuron with different tunable materials including free carriers (Si, ITO), QCSE (III-V), Pauli- and state-blocking (Graphene, exciton in TMD) [2,7] for the nonlinear activation of the photonic perceptron. With respect to noise, we include thermal and shot noise the photodetector, and capacitor charging noise of the EOM. We then derive a closed form equation set for this photonic neuron as a function of a) laser power into the PNN, and b) modulator length with signal-to-noise (SNR) ratio as an output (e.g. downstream layer of the PNN). Note, in such a PNN network, not only power needs to be considered but also the SNR for signal cascadability (i.e. Multi-layered NNs). Here the only element in the network improving SNR is the nonlinear behavior of the EOM; the steeper the transfer function the lower the MAC/J (higher efficiency), however, if it is too steep, then noise from the detector is unable to train the photonic neuron during gradient descent algorithm (in particular for higher bit-rates). To gain further insight into the dynamic behavior of such photonic neurons, we performed a MIST dataset [11] (set of images of handwritten digits in a grayscale 28x28 pixel format) inference test, after the PNN was trained (Adagrad [14] method) to classify the images into the 10 individual digits, and thence evaluated PNN accuracy (**Fig. 3**). The sigmoid activation function was replaced with a custom activation function from the EOMs, including noise (weights are bound [0,1]) to simulate input optical weighting by microring ring tunable filters [1]. The results show that quantum well and quantum dot modulation outperforms graphene and exciton modulation in terms of accuracy in the low laser power limit. As optical power is increased, the graphene and exciton modulation approaches the accuracy of quantum well neurons (**Fig. 3**). *In conclusion*, photonic neural networks are a promising class of processors able to surpass the CMOS efficiency wall and are not bound by charging electrical wires, due to their synergistic physical implementation of a perceptron model in integrated photonics.

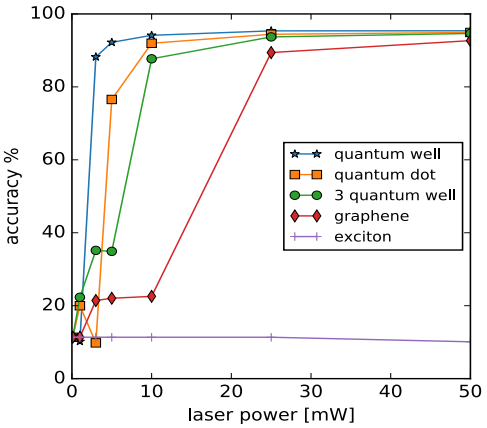

Figure 3. MNIST classification accuracy for the PNN using photodetectors ring-resonators for VMM, photodetectors for summation, and EOM (legend: different EO materials [3,7-10]) as nonlinear activation. Improved accuracy is found for higher modulation efficiency. NN: 2 layer each 100 photonic neurons; learning rate=0.005, 45 training epochs, no decay, used Keras [12] and TensorFlow [13]

References

[1] A. N. Tait, M. A. Nahmias, B. J. Shastri, et al., J. of Lightwave Technology, vol. 32, 21, pp. 3427–3439, 2014.
[2] R. Amin, J. B. Khurgin, V. J. Sorger, Opt. Exp. 26, 12, 15445-15470. 2018.
[3] Z. Ma, R. Hemnani, L. Bartels, R. Agarwal, and V. J. Sorger, Applied Physics A, vol. 124, no. 2, p. 126, 2018.
[4] R. Amin, Z. Ma, R. Maiti, S. Khan, J. B. Khurgin, H. Dalir, and V. J. Sorger, Applied Optics 57, 18 (2018).
[5] V. J. Sorger, R. Amin, J. B. Khurgin, Z. Ma, et al., J. Optics, vol. 20, no. 1, p. 014012, 2017.
[6] D. A. Miller, Journal of Lightwave Technology, vol. 35, no. 3, pp. 346–396, 2017.
[7] R. Amin, C. Suer, Z. Ma, I. Sarpkaya, J. B. Khurgin, et al. Nanophotonics, vol. 7, 2, pp. 455–472, (2017).
[8] Z. Ma, et al., IEEE Journal of Selected Topics in Quantum Electronics, vol. 23, no. 1, pp. 81–88, 2017.
[9] C. Ye, K. Liu, R. A. Soref, V. J. Sorger, Nanophotonics, 4(1), 261-268. (2015).
[10] R. Amin, C. Suer, Z. Ma, I. Sarpkaya, et al. Solid State Electronics, 136, 92-101. (2017).
[11] Y. LeCun and C. Cortes, [Online]. Available: http://yann.lecun.com/exdb/mnist (2010).
[12] F. Chollet et al., "Keras," https://keras.io, 2015.
[13] M. Abadi, et al., Software available from tensorflow.org (2015)
[14] J. Duchi, E. Hazan, Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. Technical Report UCB/EECS-2010-24, EECS Dept, UC Berkeley (2010).