All-Photonic Artificial Neural Network Processor
Via Nonlinear Optics

Jasvith Raj Basani
jasvith@umd.edu

Stefan Krastanov
stefankr@mit.edu

Mikkel Heuck
mheu@fototonik.dtu.dk

Dirk R. Englund
englund@mit.edu

Abstract: We propose an all-photonic architecture to accelerate linear matrix processing by encoding information in the amplitudes of frequency states. Our design is unique in providing a unitary, reversible mode of computation at high speeds. © 2022 The Author(s)

1. Introduction

Neural networks have enjoyed widespread success with applications ranging from language processing to game playing. Traditional digital electronic architectures, however, face a bottleneck in linear matrix processing, which is a fundamental operation in machine learning. Integrated optics has recently captured interest as a platform to accelerate linear matrix processing. In this paper, we propose a novel architecture for a completely photonic implementation of a neural network accelerator via non-linear optical intermodulation. Our architecture is unique in providing a reversible mode of computation at arbitrary large speeds, to within thermal limits posed by the circuitry.

2. Programmable Linear Transformations

We begin developing the programmable linear transformations by considering a microring resonator that supports a frequency comb. Each frequency state is chosen to be either a neuron mode or a pump mode that interact with each other via the process of Four-Wave Mixing (FWM). Information is encoded into the complex amplitude of the neuron modes which can be modulated via controlled interactions with the pump modes that encode the linear weight transformations. The Hamiltonian associated with a single instance of FWM between two pump modes (\(\hat{p}_1, \hat{p}_2\)) and two neuron modes (\(\hat{a}_1, \hat{a}_2\)) is given by

\[
\hat{H} = \chi_{\text{eff}}^{(3)} (P_1 P_2^* a_1 a_2^* ) + \text{H.C.}.
\]

The pumps are assumed to be strong classical modes of light and the coupling coefficient \(\chi_{\text{eff}}^{(3)}\) accounts for the nonlinear susceptibility of the material, mode volume of the cavity and phase matching conditions.

We propose two models for the programmable linear transformations - the methods of passive and active capture and release. In the former, we consider propagating neuron modes that are captured passively into the microring resonator. The caveat is that, however, this method can only be used in the steady state regime under the slowly-varying pulse envelope approximation. The Heisenberg-Langevin equations [1] allow us to solve for the time evolution of the neuron mode amplitudes \(A_j\) in terms of the pump mode amplitudes \(P_i\), input pulse profile \(S_{\text{in}}\), \(\chi_{\text{eff}}^{(3)}\) and total loss rate \(\Gamma\).

\[
\frac{dA_j}{dt} = \left( -\frac{\Gamma}{2} + i \chi_{\text{eff}}^{(3)} |P_1|^2 \right) A_j - \chi_{\text{eff}}^{(3)} \left( \sum_{j=1}^{N} (P_i P_j^*) A_j - \sum_{j<i} (P_i P_j^*) A_j \right) - \sqrt{\Gamma} S_{\text{in},i}
\]

Written in matrix format, we find that the relationship between the amplitudes of the input modes and the neuron modes results in a Toeplitz matrix, i.e., a matrix which is constant by diagonal. This implies that a single transformation matrix spans only a fraction of the space spanned by the set of unitary matrices. To quantify the group of operations spanned by these Toeplitz matrices, we introduce the concept of expressivity that we measure numerically using the
Fig. 1. Left: Schematic of the neural network, showing information encoded in amplitudes of frequency states. Consecutive layers are shown to be made of microring resonators, connected by a Lithium Niobate waveguide. Right: The polar plots show the nonlinear action function. The output phase (top) and output amplitude (bottom), shown in color are illustrated as functions of the input phase (azimuthal plane) and input amplitude (radial direction).

The method of passive capture and release, however, requires long gaussian pulses and a large number of microring resonators to implement a single layer of a deep neural network. The method of active capture and release overcomes this by actively capturing the neuron modes through a tunable coupling coefficient, storing them in the resonator to perform FWM with stronger time-dependent pumps. The time dynamics of the neuron modes is mathematically similar to (1) without the contribution from the input mode profile. The solution is thus an exponential of a Toeplitz matrix and presents with limited expressivity. Just as in the previous case, we find that cascading a sufficient number of FWM operations allows us to span the whole group of unitary matrices.

3. Nonlinear Activation Function

The nonlinear activation function we propose comprises of two major components. The first component of our nonlinearity relies on the interaction of the neuron modes with externally pumped subharmonic modes. This interaction is engineered to occur in a waveguide with a second order nonlinear susceptibility placed between consecutive microring resonators. This interaction can be described mathematically by a set of coupled partial differential equations [2]. The result of this nonlinear interaction is a distortion in the amplitude and phase profiles of the neuron modes.

The second component of the nonlinearity relies on the controllable capture of the neuron modes into the microring resonator. The controllable capture is enabled by the tuning the coupling coefficient $\gamma(t)$. This coupling coefficient is optimized such that modes with only a gaussian amplitude profile are captured perfectly. The distortion in the profiles of the neuron modes created by the first step results in their imperfect capture, thus enhancing the degree of nonlinearity. We illustrate the polar plot of the net nonlinearity as a function of the input phase and amplitudes in fig. (1).

4. Discussion

The scheme we propose has multiple advantages over previous implementations of optical neural networks. Primarily, since we intermodulate all the neurons in a single step, the time complexity is $O(N)$. Furthermore, the parametric nature of the operations accounts for the reversibility and allows us to compensate for losses as well. We also find that the speed of our operations is proportional to the powers of the pump modes, thus enabling large speeds at high powers. Finally, advances in experimental techniques enables the construction of high Q-valued ring resonators to realize this scheme experimentally.

References

1. Heuck, M. and Koelewoed, J. G. and Christensen, J. B. and Ding, Y. and Frandsen, L. H. and Rottwitt, K. and Oxenløwe, L. K., “Unidirectional frequency conversion in microring resonators for on-chip frequency-multiplexed single-photon sources,” New J. of Phys., 21, 033037, (2019)
2. Kenneth D. Shaw, “Nonlinear spatio-temporal theory of the doubly resonant degenerate optical parametric oscillator”, Proc. SPIE 2379, Solid State Lasers and Nonlinear Crystals (1995)