Optical readout for low resolution weighting and easy observation for integrated photonic reservoir computing

Chonghuai Ma[1], Floris Laporte[1], Stijn Sackesyn[1], Joni Dambre[2] and Peter Bienstman[1]

[1] Photonics Research group, UGent – imec, Technologiepark 126, 9052 Gent, Belgium
[2] IDLab, UGent – imec, Technologiepark 126, 9052 Gent, Belgium

Optical readout systems in photonic reservoir computing with non-volatile weighting elements will result in huge savings in power consumption. However, the low precision weighting that comes with this will have a severe impact on system performance. Our proposed method can improve the performance by one order of magnitude in terms of bit error rate. Moreover, in the optical readout system, lacking direct node observation is a problem. We present our latest design that can enable easier observation on the nodes, which will both benefit system operation and laboratory measurement.

Introduction

Photonic reservoir computing [1, 2] is a promising hardware implementation of machine learning algorithms and therefore solves a lot of problems with much larger bandwidth and lower power consumption compared to conventional Von Neumann architecture processing units. In a photonic reservoir system, a readout system is necessary to read the signal out from the neural network and apply weights on the signal. Readout systems can be one of the two types: electrical readout or optical readout [3]. An electrical readout system uses external or on-chip photodiodes to receive output optical signals from all the nodes and convert them to electrical signals. These signals are then weighted and summed in electronic microcomputing units. However, the integrated optical readout system uses optical weighting elements such as micro-ring resonators or Mach-Zehnder interferometers (MZI) to apply weights directly in the optical domain, as well as the signal summation, by using waveguide combiner tree. Compared to the electrical readout, the optical readout has a much higher bandwidth and lower power consumption. Therefore it is more compatible with the optical reservoir structure.

In an all-optical readout system, the weighting is done in the optical domain, usually by integrated heater based optical phase shifter and amplitude modulators. However, by using non-volatile weighting elements [4], no further power consumption is required once the weight has been set. In contrast, heaters require a constant power supply to maintain the weights during operation. The power consumption of the weighting can be reduced to nearly zero using non-volatile elements. The drawback is that these weighting elements will only be able to provide very limited resolutions. In deep learning, people refer to limited resolution weighting problems as 'weight quantization'. Typically, starting from the quantization of a full-precision pre-trained model, a subset of weights is selected to be kept fixed, and the other weights are then retrained in full precision and re-quantized. This step can be iterable depending on the individual model to achieve optimal performance.

In the photonic neuromorphic context however, the low-resolution weighting problem is different compared to a typical 'weight quantization' in deep learning. First, photonic reservoir computing systems usually have around a hundred nodes which, compared to deep learning models with millions of parameters, doesn't provide much redundancy for an accidentally 'wrong' selection of the retraining weight connections. Furthermore, the optical weighting elements have more restrains in terms of insertion loss and extinction ratio, and more importantly, they have severe weighting noise. The proposed method we came up with is inspired by the weighting pruning and retraining method used in the deep learning fields with a special emphasis on the weight selection method.

For an all-optical readout in a reservoir computing system, node observation problem is always a tricky part. The signals are coherent optical signals and obtaining both amplitude and phase information simultaneously is not trivial. Especially during the training phase, it is necessary to be able to measure every node state hopefully in a short time. One solution to this is called nonlinearity inversion [5]. It requires a reference signal and needs repetitive measurement over all the signals. One drawback of the method is that to measure the signal from one node, all the weighting elements from other nodes have to be involved to help to block signals from their channels. Due to the low extinction ratio or possible manufacturing deviations, the signal-to-noise ratio (SNR) of the target measurement channel can be uncontrollable. In this paper, we propose a method that works together with the weighting element only from its own channel to get individual observation at each node. The method will benefit the measurement and training of the system and enable both electrical readout schemes and optical readout schemes.

Low-resolution weighting

We target a optical weighting element model that can be used as a generic representation with three major aspects taken into account to match a realistic component: resolution, extinction ratio, and noise. For passive optical weighting elements, we set the maximum weight of the amplitude to 1. The minimum weight is $\omega_{\min}$. This is related to the extinction ratio as follows:

$$\omega_{min} = \frac{1}{extinction\ ratio}$$

In the quantization procedure, weights below $\omega_{\min}$ will be rounded up to $\omega_{\min}$. The interval between two adjacent weighting levels is given by:

$$\Delta\omega = \frac{1 - \omega_{min}}{N - 1}$$

where N is the total resolution of the weighting element.

Weighting noise varies from technology to technology. Here we simply apply an approximation of a Gaussian distribution with a standard deviation of σ. The noise coefficient we will use to represent the noise level will be given by:

$$noise\ coefficient = \frac{\sigma}{\Delta\omega}$$

The training procedure we take is following a procedure of weight quantization - weight selection - retraining. As a starting point, we take a pre-trained full-precision model, then apply a general quantization. The performance should be drastically degraded as the system resolution is extremely low (later referred as naive low precision weighting). Then we choose several random partition selections of the weights to choose which to be fixed and which to be retrained in the next retraining step. By comparing the performance of these possible partitions, we are able to choose the best one. The whole procedure can be deployed in an iterable way and progressively change the ratio of the number of weights to be fixed or retrained.

With the retraining method, readout system with limited weighting resolution can deliver a performance much better than a naive implementation of limited weighting, and perform close to high-precision weighting systems. Table 1 shows a simulation result of our method on a 4x4 4 port reservoir architecture [6] with optical weighting elements that have a low resolution of 32 levels in total, and with severe weighting noise involved. The task is 3-bit delayed XOR. The results show that even given a fairly difficult task, naive quantization weight can hardly deliver convincing performance across three different noise levels. However, our proposed retrained method is able to provide up to 1 order of magnitude better performance for given 32 levels of resolution better than naive weight quantization, and follows the high precision weights performance closely.

Noise coefficient	0.0025	0.005	0.01
High precision weighting (BER)	0.0072	0.0340	0.1073
Retrained low precision weighting (BER)	0.0150	0.0433	0.1191
Naive low precision weighting (BER)	0.1093	0.1194	0.1554

Table. 1. Performance comparison between high precision, low precision with retrained method and naïve low precision weighting resolutions in bit error rate (BER)

Easy observation readout architecture

The proposed architecture is illustrated below as figure 1. As shown, one Mach-Zehnder interferometer (MZI) is used for weighting both amplitude and phase for each node input and the whole structure shows a simple 2-node input combining as a part of a whole readout system.

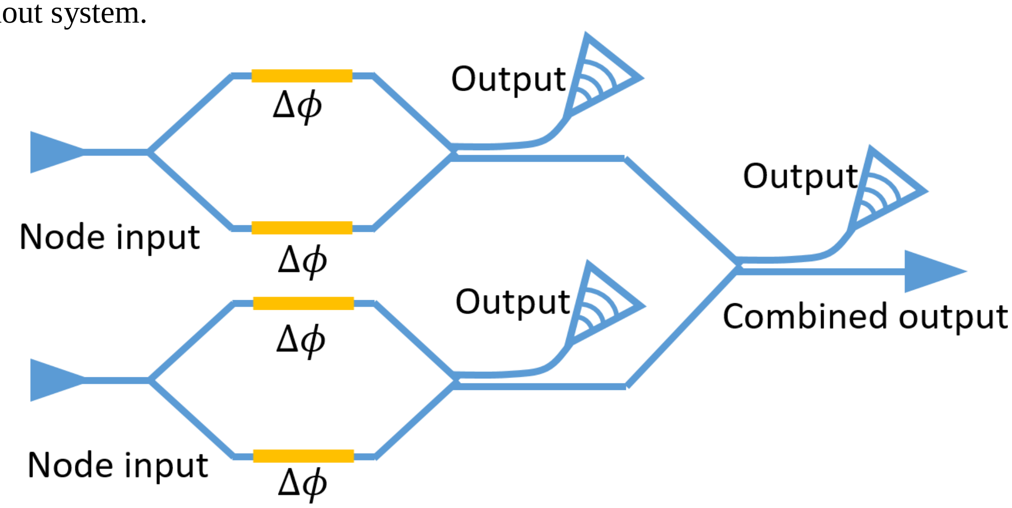

Figure. 1. Illustration of proposed easy observation readout architecture, consisting of 2 channels and one combiner.

Each node will have its own grating coupler for easy observation of the signal amplitude. The amplitude measurement is completely independent and is not affected by any other node signals. The phase observation of all the nodes can be followed up by making use of the amplitude information just obtained. Let the amplitude of the three observation points be a, b, and c, with their phase ϕ_a, ϕ_b, and ϕ_c. The last signal is dependent on two input signals on the input side of the combiner. The three signals can be written as:

$$ae^{i\phi_a} + be^{i\phi_b} = ce^{i\phi_c}$$

From the trigonometric functions we can get:

$$\cos(\mathrm{A}) = \frac{\mathrm{b}^2 + \mathrm{c}^2 - \mathrm{a}^2}{2\mathrm{bc}}$$

$$\cos(\mathrm{B}) = \frac{\mathrm{a}^2 + \mathrm{c}^2 - \mathrm{b}^2}{2\mathrm{ac}}$$

$$\cos(\mathrm{C}) = \frac{\mathrm{a}^2 + \mathrm{b}^2 - \mathrm{c}^2}{2\mathrm{ab}}$$

where A, B, and C are the angle of the triangle that consists of the three edges a, b and c. Therefore, from the calculation we know, given the phase information of combined signal ϕ_c, the phase of the two inputs can be given by:

$$\phi_b = \pm A + \phi_c$$

$$\phi_c = \phi_c \pm B$$

Until now, we can get the phase information exclusively calculated by the amplitude information we get from the measurement and the phase information from the combined signal. There are two sets of solutions to the phase ϕ_a and ϕ_b, the finalization for these can be measured by a perturbation on one of the inputs signals. Depending on how the final signal changes, the solution can be determined correspondingly.

Conclusion

Low-resolution weights are a major challenge when using non-volatile weighting elements to achieve ultra-low power consumption photonic reservoir computing systems. By implementing our proposed algorithm during training, the performance on very low weighting precision (32 levels) is much better than naive quantized weights and is very closed to full-precision weights, even taking severe weighting noise and bad extinction ration into account. On the other hand, with our proposed readout architecture, we prove the concept of a new easy node observation method for all-optical readout systems. By observation of all three (nodes and combined) signals, the phase relation between the two inputs can be calculated correspondingly with an additional perturbation measurement of the node signal. Therefore, both amplitude and phase are obtained without the potential crosstalk from other output channels.

References

[1] H. Jaeger and H. Haas, "Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication," *Science,* vol. 304, no. 5667, pp. 78-80, 2004.

[2] W. Maass, T. Natschläger, and H. Markram, "Real-Time Computing Without Stable States: A New Framework for Neural Computation Based on Perturbations," *Neural Computation,* vol. 14, no. 11, pp. 2531-2560, 2002.

[3] C. Ma, S. Sackesyn, J. Dambre, and P. Bienstman, "All-Optical Readout for Integrated Photonic Reservoir Computing," in *2019 21st International Conference on Transparent Optical Networks (ICTON)*, 2019, pp. 1-4.

[4] S. Abel *et al.*, "A strong electro-optically active lead-free ferroelectric integrated on silicon," *Nature Communications,* Article vol. 4, p. 1671, 04/09/online 2013.

[5] M. Freiberger, A. Katumba, P. Bienstman, and J. Dambre, "Training passive photonic reservoirs with integrated optical readout," *IEEE transactions on neural networks and learning systems,* 2018.

[6] S. Sackesyn, C. Ma, J. Dambre, and P. Bienstman, "An enhanced architecture for silicon photonic reservoir computing," in *Cognitive Computing 2018-Merging Concepts with Hardware*, 2018, pp. 1-2.