A silicon photonics feed-forward neural network for nonlinear distortion mitigation in an optical link

MATTIA MANCINELLI, PAOLO BETTOTTI, and LORENZO PAVESI

Nanoscience Laboratory, Department of Physics, University of Trento, via Sommarive 14, 38123 Povo (Trento), Italy. *paolo.bettotti@unitn.it

Abstract: We design and model a single-layer, passive, all-optical silicon photonics neural network to mitigate optical link nonlinearities. The network nodes are formed by silicon microring resonators whose transfer function has been experimentally measured. Both the transmitted amplitude and phase maps of the nonlinear response of the microrings are parametrized as a function of the wavelength and of the signal power to form tunable activation functions of the single nodes in the complex valued network. Training of the network is achieved by a particle swarm optimizer which selects the complex weights and the activation functions. We demonstrate that a single feed-forward layer with a single node perceptron is effective in compensating linear and nonlinear distortions over a broad range of signal-to-noise-ratio and propagation lengths. We propose to implement this simple neuronal network as an optical link transparent layer to correct signal distortions.

© 2020 Optical Society of America

1. Introduction

The capacity and reach of an optical fiber link is limited by the fiber’s optical nonlinearities (Shannon limit) [1,2]. Rising the signal power to increase the optical signal to noise rate (OSNR) causes significant nonlinear signal distortions which in turns degrade the bit-error rate (BER). Therefore, effective error correction algorithms have been developed [2]. However, the large bit rate used in modern optical communications is a challenging task for nonlinear mitigation methods based on digital signal processing [3,4]. This constitutes a significant part of the final cost and power budget of the communication link. Thus, there is the need for more effective, faster and energy saving alternatives. An all-optical implementation of signal correction based on machine learning techniques could be an effective solution [5,6]. In fact, the bandwidths of many nonlinear optical effects, useful for optical neural network (NN), are of the same order of magnitude of the bit-rate of single optical channel coding (from few to up to tens of Gbps). Thus, an all-optical implementation of a neural network could act as a transparent correction layer for the optical link without affecting its operations.

Optical NNs are made of nodes having a nonlinear transfer function and connected through a defined topology [7–9]. In integrated photonics, two main types of schemes were so far proposed: feed forward networks (FFN) or recurrent networks. These last, with particular emphasis on reservoir computing networks [10,11], exploit time dependent propagating signals within nodes admitting backward loops. Recurrence provides rich network internal states (high dimensional data representation) and supplies memory to the network about past inputs. Reservoir computing photonic integrated circuits were demonstrated [12–16] but their optimization is mostly approached heuristically. Only recently, their train has been modeled, albeit using simple, non realistic, node nonlinearities [17–19]. On the other hand, FNNs are the standard model for software based NN as they permit an efficient optimization of the network parameters that are learned during the training phase. Following the pioneering work of [20], recent
works demonstrated photonics FFNs with highly efficient matrix multiplication schemes and back-propagation training algorithms [21–24]. Yet, these networks do not implement nonlinear optical function in the integrated photonic chips, being this step confined to the detection or performed off-chip.

In this work, we model a passive photonic FNN where the nodes are described by the measured nonlinear characteristics of silicon microresonators (MRs) used in the all-pass configuration. Both the phase and the amplitude of the optical signal is considered, i.e. the NN treats complex valued quantities. Specifically, we propose to train both the complex valued weights as well as the MRs resonance frequencies, i.e. the signal power dependent phase and amplitude of the MR transfer characteristics. In addition, our realistic model of an integrated photonic NN considers energy conservation (i.e. power losses) and positive-only signal values. Thus, while software based NN are free to choose among positive and negative weights and biases for node connections, our FFN use only positive values. The lack of negative weights/biases strongly limits the space search as well as the chance to effectively train the network. This limit has been overcame by coherent schemes where interference is used to sum positive and negative field values. However in our approach we simplify even further the network structure to develop a more robust and compact device. To minimize the number of connections among nodes, whilst keeping high network computational capability, we consider a 1xNx1 FFN, i.e. an input signal which is distributed to a $N$ nodes layer coupled to a single output perceptron. The main advantages of the proposed NN are:

1. it is fully passive as it requires only standard trimming functionalities to tune the MR at the desired working points.

2. Silicon photonics MRs are used. Neither other materials (eg. phase changing materials [25]), nor active semiconductors [16], nor E/O conversion steps [26] are used.

3. The nodes are made by a MR which acts as a nonlinear filter. Since each MR is used in the all-pass configuration, its nonlinearity is provided by its phase bistable response [27,28]. This approach maximizes the nonlinear response of the NN and, allows correcting highly distorted signals while using small and passive NN.

4. The resonance tuning of the MRs provides a trainable node activation function. This differs from most of the actual neural networks where a fixed activation function, equal for all nodes, is used. This possibility comes at no cost as it is an hyperparameter optimized during the training by simply tuning the resonance of each optical node and increases the NN computation capability.

5. We use the MR phase response dependence on the resonance frequency as an artificial, positive bias. In fact, changing the MR resonance changes the node input power at which the MR bistable behavior occurs (see more in Sec. 2.1).

6. Since the optical signals encode both amplitude and phase information, their complex valued representation enriches the NN response and permits to minimize the number of nodes [29,30].

7. It works in direct detection mode, thus simplifying the circuit design.

8. A particle swarm optimization scheme allows to train the non-differentiable parameters such as the complex-valued node activation functions and to manage the hyperparamentric activation function [31].

9. The real and imaginary parts of the complex-valued weights are trained independently.

In this work, we validate the proposed NN as a transparent layer to recover the information for nonlinearly distorted signals in a noisy optical link.
2. Physical model and network

Fig. 1. Map of the phase delay $\phi$ between the output field $S_{\text{out}}$ and the input field $S_{\text{in}}$ of a microring as a function of the detuning and $P$. The colors in the map are simply guides for the eyes. The inset shows the all-pass configuration of a microring resonator.

2.1. Microring resonator as nonlinear optical nodes

The activation functions $f$ of the trainable nonlinear nodes of our NN link the node input signal ($S_{\text{in}}$) to the node output signal ($S_{\text{out}}$). The $i$-th node activation is a complex function $f_i(\Delta \lambda, P) = S_{\text{out}}^i / S_{\text{in}}^i = A_i \exp(j \phi_i)$, with $\Delta \lambda$ the wavelength detuning, $P = |S_{\text{in}}|^2$ the input power, $A_i$ and $\phi_i$ the amplitude and the phase of the field transmitted by the node. The MR response is assumed to be instantaneous with respect to the variation of the input signal.

The activation functions are based on the experimental measured transmission spectra of silicon microring resonators under different $P$. To build the matrix of $A_i$ and $\phi_i$ values, parameterized for different $\Delta \lambda$ with respect to the cold resonance wavelength and $P$, we used data from silicon microrings fabricated on SOI (silicon on insulator) wafers. The geometric parameters of the MR
were waveguide cross section $450 \times 220$ nm, ring radius $7 \mu m$, bus waveguide-ring coupling $\sim 3\%$, ring Q-factor $\sim 10^4$ and cold resonance wavelength $1538.74$ nm. Measurements were performed on microrings in an add-drop configuration. Then, the phase response was estimated from the theoretical fit of the measured drop-port and through-port transmissions. Fig. 1 shows the parameterized map for $\phi(\Delta \lambda, P)$.

Without losing generality, the microring transmission spectra were measured in the (slow) thermal nonlinear regime. Faster nonlinear responses, such as the free carrier induced nonlinearity \cite{27,28} or the third order Kerr nonlinearity \cite{32}, could be exploited as well. Care has to be taken in the excitation of the node in order not to enter the self-pulsing or the chaotic regimes which makes the NN untrainable \cite{33}.

Despite the add-drop configuration shows the bistable behavior is observed both in the amplitude and in the phase response of the node activation function \cite{34,35}, we develop our NN based on microring in the all-pass configuration. This because the device losses in the all-pass configuration are significantly smaller that in the add-drop one when using the drop transmission. Simulations show that these losses spoil the computation capacity of the purely passive NN due to the significant signal attenuation after each node, even for small $\Delta \lambda$. Remarkably, we will show that the bistable behavior of the phase response only is enough to achieve quite accurate NN performances. Since the phase discontinuity threshold depends on $\Delta \lambda$, we exploit this fact as an artificial, positive bias to tune the point at which the phase jumps from the low to the high state with respect to the input signal wavelength. In addition, in the all-pass configuration the amplitude and the phase are disentangled in the node activation function. Therefore, the two can be addressed separately during the training phase which allows regulating separately the sum amplitude of the real and imaginary parts of the complex value weights.

2.2. Photonic Neural Network

Since any optical signal is described in the complex plane by its amplitude and phase, the NN deals with complex-valued quantities (e.g. input sequences, node activation functions and weights). The feed-forward network topology is quite simple: the input sequence is distributed to a layer of $N$ nodes whose outputs are connected to a single node complex perceptron \cite{36}. The perceptron output is then detected (i.e. the output signal is squared since detection is performed by a photo-detector which is sensible to the output signal power $P_{out}$) and passed through a discriminator, which based on a trainable threshold, outputs the binary sequence. This is used to compute the NN performance (i.e. the BER). On the other hand, the training of the NN is performed by reducing a loss function $L$ computed on the output complex signal from the perceptron (e.g. as in a standard classification task). This simple NN topology is schematically shown in Figure 2(c). More formally the NN acts as

$$X = f \odot B$$

(1)

where $X$ is the vector of the layer state (i.e. the nodes output signals $S_{out}^i$, $i=1, ..., N$), $f$ is the matrix of the activation functions selected by the optimizer, $B$ is the vector of the h-inputs of the sampled portion of bits (see more later) and $\odot$ is the element-wise product. The output perceptron acts as:

$$X_o = f_0(WX) = \left( \sum_i w_i e^{j\varphi_i} S_{out}^i \right) e^{j\theta_0}$$

(2)

where $X_o$ is a output complex value, $W$ is the vector of weights ($w_i, \varphi_i, i=1, ..., N$) applied to $X$ and $f_0(\cdot)$ is the activation function of the output node.

The loss function $L$ is assumed to be the squared sum of the residuals of the real and imaginary parts:

$$L = \sum \left[ (\Re(T) - (\Re(X_o))^2 + (\Im(T) - (\Im(X_o))^2 \right]$$

(3)
where $T$ is the target sequence (i.e. the complex valued signal sequence which has not been
distorted by the propagation in the optical fiber link). $\mathbb{R}$ and $\mathbb{I}$ are, respectively, the real and
imaginary parts.

2.3. Optimizer

The training (optimization) of the passive photonic neural network aims at reducing $L$ by
changing the trainable parameters while fulfilling the following physical constrains:

$$ A_i, w_i \leq [0, 1] $$ (4)

$$ \phi_i, \varphi_i \leq [0, 2\pi] $$ (5)

$$ P_{in} \leq P_{out} $$ (6)

that is: weights are non negative and at most equal to unity, the phase of each node is limited
between 0 and $2\pi$ and, finally, the power entering the NN can only be reduced at its output ($P_{in}$
is the average power at the input of the optical fiber link).

Since the phase response is a non-differentiable parametric function and the NN is complex-valued,
the use of a standard back-propagation algorithm is heavily hindered [37]. In fact, given
the impossibility to have a bounded and analytical functions over the complete complex plane (as
for the Liouville’s theorem), the training requires ad-hoc activation function that permit to train the
NN by using their complex-valued nature. To overcome this, we used a Particle Swarm Optimizer
(PSO) [31,38]. PSO has two advantages: 1) it is not limited to differentiable functions and 2) it
easily permits to handle the parametric form of the node activation functions. In addition, the
training has to split the complex values into real and imaginary parts that are concurrently trained.
Albeit the latter approach does not fully exploit the richness of complex-valued description,
since the splitting and independent training of the real and imaginary parts removes the link
between them, it permits to use the activation functions we derived in Sec. 2.1. As we will
show later-on, training the activation functions further enriches the NN capability to project the
original problem into higher dimensional space and ease its separability. The drawback of the
PSO is its stochastic nature that does not guarantee reaching the global minimum for problem
involving a large number of parameters. This fact, together with the considerations about energy
dissipation and integration level, motivates the single layer topology. In this way, the number
of nodes is limited, which in turn reduces the number of trainable parameters so that the PSO
algorithm converges to a pseudo-global minimum in a repeatable and rapid way.

The PSO creates a population of states (the $\theta$ vector) and move each individual (i.e. each
vector element) over the phase space to minimize the loss function. The parameter $\theta$ vector
encoding all the system variables to be optimized is $\theta = \{\{w_i, \varphi_i\}, \{f_i\}\}$, where:

- $\{w_i, \varphi_i\}$ are the weights (subscripts $i$ label the node, $i=1, \ldots, N$).
- $f_i$ is the specific activation function associated with each neuron ($i = 1, \ldots, N$ and $i = 0$
  for the output node).

Once training is done, the optimal parameter vector $\theta^*$ is used at inference time to test the NN
performances.

2.4. Optical Link Model

The optical link we simulated is sketched in Fig.2(a) [39]. A $2^8$ bits-long Pseudo Random
Binary Sequence (PRBS) is generated, sampled at 26 samples/bit, low-pass filtered to simulate
the optical link bandwidth (with the threshold at 0.7 the bit-rate of 10 Gbps), pre-compensated
(leaving 700 ps/nm of residual delay), degraded by adding a random noise up to the desired OSNR and propagated up to the chosen fiber span (1 - 100 km). During the propagation, both linear (eg. dispersion) and nonlinear (eg. four wave mixing and self phase modulation) effects distort the original bit sequence.

Fig. 2b shows the signal degradation after propagating a fiber span of 89 km, with a $P_{in}$ of 8 dBm in presence of an OSNR of 12 dB. This waveform represents the signal to be treated by the NN. Our modeling assumes a one-bit inter-symbol interference. To cope with this, the NN input is sampled from a time window $tw$ spanning three bits. Within $tw$, each bit is sampled at 26 equally-spaced times (for clarity, only 6 sampling time-slots are shown in Fig. 2(c)). These $26 \times 3 = 78$ data form the input vector to the NN. The NN divides the input vector (of elements $x_n, n = 1, \ldots, 78$) in $h=78/N$ subsets each with $N$ data and spaced by $h$ (the first sub-set has $x_1, x_{h+1}, x_{2h+1}, \ldots$, the second sub-set $x_2, x_{h+2}, x_{2h+2}, \ldots$, and so on) and inputs each subset to the $N$ nodes. The NN output is stored. Then, this process is repeated on the other sub-sets. Therefore, the NN processes all the data from the three sampled bits. Finally, the NN minimizes $L$ by comparing each output to the first bit in $tw$ and searching for the optimum parameter set. Referring to Fig. 2c, the first bit is labeled $(n + 2) - th$. The process is re-iterated over the entire bit sequence.

Since the PRBS contains all the possible combinations of $2^8$-bit strings, the training phase is fast as it requires only a few tens of noisy PRBS (see also Fig.3). On the other hand, the testing is done by using 5,000 noisy PRBS to obtain relevant statistics at low BER values. BER is calculated using the error counting method [39]. Note that both the optimal sampling threshold as well as the optimal sampling time among the 26 samples per bit are also optimized during the training.

The overall insertion loss of the NN depends on the actual network configuration. It can be roughly estimated as the sum of the coupling losses to the NN chip ($2 \times 1$dB) and to the device losses of two microring resonators (the one in the layer and the output one) which typically is 5 dB.

3. Results

The performance of the NN was evaluated by comparing the BER of the same fiber-link without the NN (reference link) and with the NN placed before the detector (for short named network). To assure that PSO converges to pseudo-global minima, we repeated the training at least three times for each NN optimal configuration. The results are reported in Fig.3. Convergence tests have been performed over two input powers (10 and 14 dBm) and OSNR (6 and 12 dB) values for a fiber span of 89 km. This tests the NN on bits distorted by significant nonlinearities. We noted that for a OSNR of 12 dB the reference link fails to recognize the bits (BER $leq$ 1), thus we use this OSNR value to train the NN. Fig.3(a) shows that networks with more than 15 nodes yield a constant $L$, whose value depends only on the OSNR but not on $P_{in}$. $N=15$ is the optimum number of nodes that yields input separability by the MR nonlinearities. Fig.3(b) shows that $L$ normalized to the number of training sequences is a constant. This is expected since each $2^8$ PRBS sequence contains all the possible combinations of bits. Thus, the training accuracy does not increase significantly with the number of sequences. Finally, Fig.3(c) shows that $L$ is nearly independent of the number of individuals used in PSO. Specifically, with more than 20 individuals only a reduced $L$ standard deviation is observed. These results demonstrate that PSO systematically converges to similar minima, albeit we cannot assure that it is the system global minimum. Training takes about 6 min on i7-4500U laptop with 8 Gb of RAM when the convergence tolerance of PSO during the training is fixed to $10^{-6}$. Note that this value is also used for all the following calculations.

When network configurations with more layers were investigated, PSO fails to converge to a
Fig. 2. (a) Block diagram of the fiber optical link. For an explanation of the different blocks see the text. (b) Distorted temporal shape of the optical signal (blue dotted-line) at the output of the optical fiber before the white noise is added. The gray and white blocks indicate the binary sequence of 1 and 0 bits, respectively. $t_{w_1}, t_{w_2}, t_{w_3}$ label a sequence of three intervals which are sequentially processed by the neural network, starting from $t_{w_3}$. (c) The actual data processing sequence to handle the input data to the nodes. Each one of the three bits (labeled $n-th, (n+1)-th, (n+2)-th$) is sampled in 26 time intervals (only 6 samples per bit are shown in the figure for simplicity). These are divided in sub-sets and cyclically fed to the nodes. The inset shows the neural network topology (one layer and the output complex perceptron). Both the weights and the node activation functions are trained. The neural network output is then discriminated (orange arrow) and the binary corrected sequence is generated.
Fig. 3. Test of the PSO convergence. In all panels, the data symbols refer to the loss functions obtained for the different input power $P_{in}$ and OSNR reported in the inset table. A fiber span of 89 km is used. (a) Loss function $L$ vs number of nodes for 30 individuals in the PSO and 15 training sequences; (b) $L$ normalized over the number of PRBS sequences used for the training vs the number of training sequence. The nodes were 15 and the individuals 30. (c) $L$ vs the number of individuals in the PSO for $N=15$ nodes and 15 training sequences.

Fig. 4. (a) BER vs OSNR for the reference link (crossed line) and the network (circled line). The red dashed line shows the pre-FEC limit. The neural network has $N=15$ nodes. A fiber span of 67 km and an input power $P_{in}=13$ dBm were used. $G$ indicates the OSNR gain due to the use of the neural network. (b) BER vs OSNR for the reference link (crossed line) and the network (dotted line). The red dashed line shows the pre-FEC limit. The neural network has $N=15$ nodes. A fiber span of 89 km and a $P_{in}=4$ dBm were used. (c) 2D map of $G$ at the pre-FEC limit. See text for a description of the grays shaded regions.
single minimum and different training experiments produced largely different $\theta$ (results not shown). We suggest that it is due to the number of parameters. Similarly, the accuracy of the network decreased by increasing the depth of the NN -e.g. by halving the numbers of nodes per layer and using two layers- thus we limited our model to a single layer NN.

The results of testing a trained NN of $N=15$ nodes and with a PSO with 30 individuals are summarized in Fig.4. The figure results from the detailed analysis of individual BER curves vs OSNR and $P_{in}$ reported in Fig. 5 of the Supplementary Information. The first left panel shows a comparison of the BER for the reference link (crossed line) and the network (circled line) for a fiber span of 67 km and a $P_{in}=13$ dBm. Due to the large nonlinearities, the reference link never shows a BER smaller than the pre-forward error correction limit (pre-FEC limit) of $10^{-3}$, while the NN is able to mitigate the nonlinearities and recovers a BER lower than the pre-FEC limit as soon as the OSNR is larger than 10 dB. The central panel shows the same comparison for a longer fiber span and a lower $P_{in}$, i.e. when only chromatic dispersion matters and the nonlinear distortions in the fiber are negligible. Here, the NN allows to decrease the value of the OSNR at the pre-FEC limit by -5.5 dB. To globally show the NN performance with respect to the reference link, we introduce the NN gain $G = OSNR^{N\text{NN}}_{\text{FEC}} - OSNR^{\text{REF}}_{\text{FEC}}$. This is defined as the difference in the OSNR required to achieve the pre-FEC limit by the network ($OSNR^{N\text{NN}}_{\text{FEC}}$) and by the reference link ($OSNR^{\text{REF}}_{\text{FEC}}$). Each BER point is the mean over three testings that differ for the random noise added to the PRBS. The standard deviation among the three is typically smaller than 10% of the BER, if significant statistic is accumulated (e.g. BER $\leq 10^{-5}$). The actual BER standard deviations are shown in Fig. 6.

The map clearly shows a consistent and systematic improvement of the BER due to the network which validates the mitigation effect of the NN on the nonlinear distortion. As expected the maximum gain $G$ is for the longest span and for the largest $P_{in}$, where nonlinear effects are the most important. The light gray shaded area indicates the regions where the reference-link does not reach the pre-FEC limit, while the dark gray shaded area is the region where our NN systematically achieves BERs $< 10^{-6}$. Since the limited dimension of the testing set ($\sim 5000 \times 255 = 1.23 \times 10^6$), errors on the BER are too significant to report reliable data. Still the network allows reaching very small BERs. To note that, since the PRBS were precompensated, both the network and the reference link show poor performances at the shortest lengths (< 45 km), see also Fig. S1. As expected, the region where the network overcomes the reference link starts at an OSNR $\sim 12$ dB, that is the value at which the NN was trained. As shown in Fig. 3(a), $L$ increases at small OSNR, thus the network performs better than the reference-link, but its overall accuracy decreases.

4. Conclusions

The proposed photonic neural network based on passive silicon microresonators shows remarkably mitigation of the signal degradation due to nonlinearities in an optical link. The single layer structure of the feed forward neural network allows the use of a limited number of microresonators which reduces the NN insertion losses which is one of the main limitations of the use of passive neural networks. The key ingredient in our NN is the training of the activation functions of the single nodes, and, specifically, the use of the MR bistable phase response as a resources in the NN training. We do also show that a single nonlinear node is a computational resource even though it is not a perceptron. This differs from what is usually assumed in other photonics implementations [16]. Actually, the realization of such a NN can be simply achieved by using thermal heaters for the MR tuning while the weights can be applied to the inputs of the complex perceptron by simple thermal phase shifters (for $\varphi_i$) and tunable Mach-Zender interferometers (for $w_i$). Both are standard components in silicon photonics. We have also shown that after a simple and fast training phase, the network reduces the BER penalty by several orders of magnitudes compared to the reference link. PRBS sequences are corrected to BER $\leq 10^{-3}$ across
broad ranges of OSNR (12 – 20 dB) and input power levels (5 – 14 dBm). It is worth noticing that the NN training is performed before the detection, which implies that the NN can be used also as a mitigation stage along the fiber with the signal kept in the optical domain without the need of an optical/electrical conversion. Our results demonstrate that a complex-valued single layer passive silicon photonics feed-forward neural network can act as a transparent mitigation stage in optical communication links.

**Funding**

**Acknowledgments**

We acknowledge NVIDIA Corporation for the donation of GPUs board through the GPU Grant Program. We thanks E. Sangineto for help in outlining the MATLAB code and Dr. M. Borghi for fruitful discussion. This work was partially supported by the European Research Council (ERC) (Grant agreement No.788793-BACKUP) and the Ministry of Education, University and Research (PRIN 2017-PELM project).

**Disclosures**

The authors declare no conflicts of interest.

**References**

1. C. E. Shannon, “A mathematical theory of communication,” The Bell system technical journal 27, 379–423 (1948).
2. P. Bayvel, R. Maher, T. Xu, G. Liga, N. A. Shevchenko, D. Lavery, A. Alvarado, and R. I. Killey, “Maximizing the optical network capacity,” Philos. Transactions Royal Soc. A: Math. Phys. Eng. Sci. 374, 20140440 (2016).
3. M. S. Faruk and S. J. Savory, “Digital signal processing for coherent transceivers employing multilevel formats,” J. Light. Technol. 35, 1125–1141 (2017).
4. K. Kikuchi, “Fundamentals of coherent optical fiber communications,” J. Light. Technol. 34, 157–179 (2015).
5. J. Mata, I. de Miguel, R. J. Duran, N. Merayo, S. K. Singh, A. Jukan, and M. Chamaan, “Artificial intelligence (ai) methods in optical networks: A comprehensive survey,” Opt. switching networking 28, 43–57 (2018).
6. F. Musumeci, C. Rottondi, A. Nag, I. Macaluso, D. Zibar, M. Ruffini, and M. Tornatore, “An overview on application of machine learning techniques in optical networks,” IEEE Commun. Surv. Tutor. 21, 1383–1408 (2019).
7. X. Sui, Q. Wu, J. Liu, Q. Chen, and G. Gu, “A review of optical neural networks,” IEEE Access 8, 70773–70783 (2020).
8. A. Lugnan, A. Katumba, F. Laporte, M. Freiberger, S. Sackesyn, C. Ma, E. Gooskens, J. Dambre, and P. Bienstman, “Photonic neuromorphic information processing and reservoir computing,” APL Photonics 5, 020901 (2020).
9. T. F. d. Lima, A. N. Tait, A. Mehrabian, M. A. Nahmias, C. Huang, H.-T. Peng, B. A. Marquez, M. Mischuglio, T. El-Ghazawi, V. J. Sorger, B. J. Shastri, and P. R. Prucnal, “Primer on silicon neuromorphic photonic processors: architecture and compiler,” Nanophotonics 9 (2020).
10. G. V. der Sande, D. Brunner, and M. C. Soriano, “Advances in photonic reservoir computing,” Nanophotonics 6, 561–576 (2017).
11. G. Tanaka, T. Yamane, J. B. Héroux, R. Nakane, N. Kanazawa, S. Takeda, H. Numata, D. Nakano, and A. Hirose, “Recent advances in physical reservoir computing: A review,” Neural Networks 115, 100–123 (2019).
12. A. Dejonckheere, F. Duport, A. Smerieri, L. Fang, J.-L. Oudar, M. Haelterman, and S. Massar, “All-optical reservoir computer based on saturation of absorption,” Opt. Express, OE 22, 10868–10881 (2014).
13. K. Vandooorne, P. Mechet, T. V. Vaerenbergh, M. Fiers, G. Morthier, D. Verstraeten, B. Schrauwen, J. Dambre, and P. Bienstman, “Experimental demonstration of reservoir computing on a silicon photonics chip,” Nat. Commun. 5, 1–6 (2014).
14. Y. Paquot, F. Duport, A. Smerieri, J. Dambre, B. Schrauwen, M. Haelterman, and S. Massar, “Optoelectronic reservoir computing,” Sci Rep 2, 1–6 (2012).
15. A. Argyris, J. Baeno, and J. Fischer, “Photonic machine learning implementation for signal recovery in optical communications,” Sci. Reports 8, 1–13 (2018).
16. A. Katumba, X. Yin, J. Dambre, and P. Bienstman, “A ag silicon photonics nonlinear equalizer for optical communications with intensity modulation and direct detection,” J. Light. Technol. 37, 2232–2239 (2019).
17. H. F. Song, G. R. Yang, and X.-J. Wang, “Training excitatory-inhibitory recurrent neural networks for cognitive tasks: A simple and flexible framework,” PLOS Comput. Biol. 12, e1004792 (2016).
18. F. Laporte, J. Dambre, and P. Bienstman, “Highly parallel simulation and optimization of photonic circuits in time and frequency domain based on the deep-learning framework PyTorch,” Sci Rep 9, 1–9 (2019).
19. H. Hazan, D. I. Saunders, H. Khan, D. Patel, D. T. Sanghavi, H. T. Siegelmann, and R. Kozma, “Bindsnet: A machine learning-oriented spiking neural networks library in python,” Front. neuroinformatics 12, 89 (2018).
20. M. Reck, A. Zeilinger, H. Bernstein, and P. Bertani, “Experimental realization of any discrete unitary operator,” Phys. Rev. Lett. 73, 58–61 (1994).
21. Y. Shen, N. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and M. Soljačić, “Deep learning with coherent nanophotonic circuits,” Nat. Photonics 11, 441–446 (2017).
22. T. Hughes, M. Minkov, Y. Shi, and S. Fan, “Training of photonic neural networks through in situ backpropagation and gradient measurement,” Optica 5, 864–871 (2018).
23. S. Pai, B. Bartlett, O. Solgaard, and D. A. B. Miller, “Matrix optimization on universal unitary photonic devices,” Phys. Rev. Appl. 11, 064044 (2019).
24. M. Y.-S. Fang, S. Manipatruni, C. Wierzynski, A. Khosrowshahi, and M. R. DeWeese, “Design of optical neural networks with component imprecisions,” Opt. Express 27, 2046–2048 (2007).
25. H. Michel and A. Awwal, “Artificial neural networks using complex numbers and phase encoded weights,” Appl. Opt. 49, B71–B82 (2010).
26. Q. Xu and M. Lipson, “Carrier-induced optical bistability in silicon ring resonators,” Opt. Lett. 31, 341–343 (2006).
27. V. K. Ojha, A. Abraham, and V. Snášel, “Metaheuristic design of feedforward neural networks: A review of two decades of research,” Eng. Appl. Artif. Intell. 60, 97–116 (2017).
28. G. P. Agrawal, Fiber-optic communication systems, vol. 222 (John Wiley & Sons, 2012).
Electronic Supplementary Information

Fig. 5. BER as a function of the OSNR for the reference link (crossed lines) and the network (circled lines) for various fiber length (L) reported on the top of each panel. The black dashed line is the pre-FEC limit. The table shows the color code used. The NN is formed by $N=15$ nodes. At the shortest link length (12 km), neither the network nor the reference link are able to correct the (mostly) linear chromatic effect (i.e. yield a BER value below the pre-FEC limit). This is due to the use of the pre-compensation in the fiber link. At intermediate lengths (45 and 67 km), signal pre-compensation allows the reference link to work at its best performance. Even in this situation, the network systematically outperforms the reference link. At the longest fiber spans, the network shows the most important performances. Only at the highest input power (14 dBm) fails to achieve the pre-FEC limit.
Fig. 6. 2D maps of the standard deviation (in % of the BER) computed by three different testing of the network differing for the actual white noise added to the PRBS. The y-axis show the signal OSNR, while the x-axis refers to different combinations of power/fiber length length: the top labels indicate the fiber length. For each length, 8 different input powers (0, 4, 9, 10, 11, 12, 13 and 14 dBm) were used. The regions with a large standard deviation are those with a poor statistics over the BER. The white regions refer to regions where the standard deviation estimate was meaningless.
Fig. 7. 2D maps of the surface of the OSNR at the pre-FEC limit for (left) the network and (right) the reference link. The white regions in the bottom part of the maps correspond to the shortest links where both the network and the reference link are unable to reach the target BER of $10^{-3}$. 