

Performance and Complexity Analysis of Bi-Directional Recurrent Neural Network Models Versus Volterra Nonlinear Equalizers in Digital Coherent Systems

Stavros Deligiannidis , Charis Mesaritakis , and Adonis Bogris , *Senior Member, OSA*

Abstract—We investigate the complexity and performance of recurrent neural network (RNN) models as post-processing units for the compensation of fibre nonlinearities in digital coherent systems carrying polarization multiplexed 16-QAM and 32-QAM signals. We evaluate three bi-directional RNN models, namely the bi-LSTM, bi-GRU and bi-Vanilla-RNN and show that all of them are promising nonlinearity compensators especially in dispersion unmanaged systems. As far as inference is concerned, our simulations show that the three models provide similar compensation performance, therefore, in real-life systems, the simplest scheme based on Vanilla-RNN units should be preferred. We compare bi-Vanilla-RNN in its many-to-many form with Volterra nonlinear equalizers and exhibit its superiority both in terms of performance and complexity, thus highlighting that RNN processing is a very promising pathway for the upgrade of long-haul optical communication systems utilizing coherent detection.

Index Terms—Digital coherent systems, fibre nonlinear optics, nonlinear signal processing, optical fibre dispersion, recurrent neural networks.

I. INTRODUCTION

OPTICAL communication links are the main highways for the exchange of trillions of data around the globe every day. The ever-increasing deployment of next generation mobile communication systems and cloud/edge computing infrastructures pushes the limits related to the required bandwidth capacity of optical communication links and enhances the need for ultra-high speed long-haul transmission. The optical communication society has launched a variety of techniques so as to deal with this demand. Current solutions involve a combination of advanced

modulation formats [1] with the use of space division multiplexing [2] and bandwidth extension towards other bands such as O-band [3]. Regardless the capacity enhancement method that is adopted, the major limitation factor of capacity will eventually be the nonlinear Shannon capacity limit of transmitted information [4], [5]. In long-haul high bandwidth optical networks, this limit is mainly attributed to Kerr-induced fibre nonlinearities and their interaction with amplified spontaneous emission noise from cascaded optical amplifiers [5].

Principal techniques that deal with nonlinearity compensation are mid-span optical phase conjugation (OPC) [4], [6], digital back-propagation (DBP) [7], and inverse-Volterra series-transfer function (IVSTF) [8]. OPC is a purely analog and thus ultra-fast technique, however it needs extra hardware such as low-noise and broadband wavelength converters and poses limitations in the transmission link topology as phase conjugation must take place at specific points with respect to the total link of a lightpath [6]. DBP is one of the most suitable post-processing techniques, proper for the treatment of both linear and nonlinear deterministic effects, since it emulates almost perfectly fiber channel through split-step Fourier with the exception of signal-noise interactions and polarization mode dispersion; however, its real-life implementation still remains impractical due to its high complexity especially when DBP is used to emulate and thus mitigate the effects of a multi-channel transmission scenario [9]. IVSTF is a less complex variant compared to DBP, however in principle it is more appropriate for mitigating intra-channel nonlinearity [8], [10]. Nonlinear Fourier transform is a promising alternative currently being investigated in the community [11], [12]. Lately, there exists an increasing interest in the investigation of machine learning techniques for the mitigation of transmission impairments [13]. Different paradigms based on artificial neural networks (ANNs) [14], convolutional neural networks (CNNs) [15], recurrent neural networks (RNNs) [16] are among the techniques that have been successfully applied mostly in intensity modulation/direct detection systems (IM/DD) and in orthogonal frequency division multiplexing (OFDM) [13].

Very recently, we proposed for the first time the utilization of a Long Short-Term Memory (LSTM) network, which is a well-known efficient RNN model for the compensation of fibre nonlinearities in digital coherent systems for multi-channel

Manuscript received March 3, 2021; revised April 26, 2021 and June 10, 2021; accepted June 22, 2021. Date of publication June 25, 2021; date of current version September 18, 2021. This work was supported by the H2020 project NEoteRIC under Grant 871330. (Corresponding author: Adonis Bogris.)

Stavros Deligiannidis and Adonis Bogris are with the Department of Informatics and Computer Engineering, University of West Attica, 12243 Egaleo, Greece (e-mail: sdeligiannid@uniwa.gr; abogris@uniwa.gr).

Charis Mesaritakis is with the Department of Information and Communication Systems Engineering, University of the Aegean, Karlovassi 83200, Samos, Greece (e-mail: cmesar@aegean.gr).

Color versions of one or more figures in this article are available at <https://doi.org/10.1109/JLT.2021.3092415>.

Digital Object Identifier 10.1109/JLT.2021.3092415

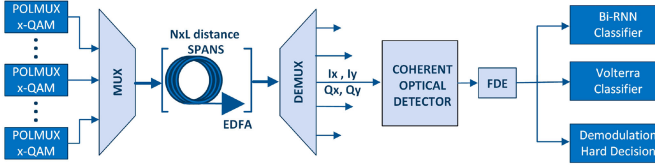


Fig. 1. The simulated transmission link.

polarization multiplexed 16-QAM systems [17]. A detailed analysis regarding the effect of LSTM model parameters and channel memory was conducted in order to reveal the properties of LSTM based receiver with respect to performance and complexity in comparison to Digital Back Propagation (DBP), proving its mitigation superiority in inter-channel transmission effects. In [18] we extended the analysis by considering two additional RNN models that are in principle less complex than LSTM, that is Gate Recurrent Unit (GRU) and Vanilla RNN in order to investigate the potential of adopting bidirectional RNN models in next generation digital coherent optical communication systems at moderate complexity. Very recently, the work of [19] confirmed experimentally the performance of the proposed RNNs in single channel operation and M. Schaedler et al. proposed a recurrent neural network soft-demapper in 800G-DWDM-600 km [20] and a soft Deep Neural Network for short reach optical communications [21].

In this work, we provide a comparative analysis of the three RNN models, namely LSTM, GRU and Vanilla RNN in WDM systems with small channel spacing and high-order modulation formats. Moreover, we compare the bi-RNN models with a 3rd order Volterra nonlinear equalizer, as a reference technique with the ability to treat complex time-dispersive nonlinear effects [22], [23], both in terms of performance and complexity. We clearly show that with the use of many-to-many training, we drastically reduce the complexity of RNN models vs. Volterra whilst keeping their performance superiority in terms of the bit-error rate (BER) of the decoded signal. The results of the present work clearly reveal the potential of bi-RNN nonlinear signal processing in next generation optical coherent communication systems as efficient and low complexity post processors. Next section describes system modelling in detail.

II. SYSTEM MODELING

The system depicted in Fig. 1 simulates the up to 1000 km fiber transmission, numerically simulated with the integration of Nonlinear Schrodinger equation (NLSE). Fibre propagation was modelled based on Manakov's equations [24] using split-step Fourier method.

$$\frac{\partial E_{x,y}}{\partial z} = -\frac{a}{2} E_{x,y} + \frac{j\beta_2}{2} \frac{\partial^2 E_{x,y}}{\partial t^2} - j\gamma \frac{8}{9} (|E_x|^2 + |E_y|^2) E_{x,y} \quad (1)$$

We consider 9-channel dense wavelength division multiplexing (DWDM) transmission with polarization multiplexing. In (1), $E_{x,y}$ contain the overall field of the nine co-propagating wavelengths including their frequency detuning with respect to

TABLE I
NUMERICAL MODEL PARAMETERS

Symbol	Parameter	Value
G	gain of amplifier	10 dB @ 50 km or 16 dB @ 80 km
a	attenuation	0.2 dB/km
β_2	second order dispersion	-21 ps ² /km
γ	fibre nonlinear coefficient	1.3 W ⁻¹ km ⁻¹
R	symbol rate / channel	25 Gbaud, 32 Gbaud, 64 Gbaud
M	modulation format	Dual-polarization 32-QAM or 16-QAM
L	span distance	50 km or 80 km
Δf	Channel spacing	35 GHz or 75 GHz @ 64 Gbaud

TABLE II
A. COMPUTATIONAL COMPLEXITY IN TERMS OF THE NUMBER OF REAL MULTIPLICATIONS – CENTRAL SYMBOL AS OUTPUT (16QAM)

Equalizer	Memory Length(symbols)	Hidden units	Multiplications
<i>Bi-LSTM</i>	151	16	386688
<i>Bi-GRU</i>	151	16	290048
<i>Bi-Vanilla-RNN</i>	151	16	96768
Volterra	151/51/11	-	14644

Computational Complexity in Terms of the Number of Real Multiplications – CENTRAL SYMBOL AS OUTPUT (16QAM)

Equalizer	Memory Length	Hid. units	Output symbols	Multiplications per symbol
<i>Bi-LSTM</i>	151	16	80	5074
<i>Bi-LSTM</i>	151	16	120	3382
<i>Bi-GRU</i>	151	16	80	3866
<i>Bi-GRU</i>	151	16	120	2577
<i>Bi-Vanilla-RNN</i>	151	16	80	1450
<i>Bi-Vanilla-RNN</i>	151	16	120	966
Volterra	151/51/11	-	1	14644

the central wavelength, thus taking into account all inter-channel nonlinear effects. We calculate the bit error rate (BER) in all scenarios studied and we aim at achieving $\text{BER} < 3.8 \times 10^{-3}$ as dictated by the hard-decision forward error correction (HD-FEC) which adds a 7% overhead. We test 16-QAM 32 Gbaud per polarization which corresponds to 256 Gb/s line rate (238 Gb/s net data rate) per wavelength and then we extend the analysis to 32-QAM at 25 Gbaud (250 Gb/s line rate and 232.5 Gb/s net data rate) and 16QAM at 64 Gbaud (512 Gb/s line rate, 476 Gb/s net data rate) so as to have a variety of modulation formats and baud rates. Net data rate is derived considering that only FEC overhead is taken into account. Lumped amplification was used with span length equal to 80 km or 50 km and noise figure equal to 5 dB. All parameters are summarized and provided in Table I. The optical receiver depicted in Fig. 1 consists of an optical hybrid, balanced photodetectors, low-pass electrical filters with cut-off frequency matched to the baud rate. We assumed ideal carrier phase and frequency estimation as well as polarization demultiplexing as we want to solely focus on nonlinear impairments and their mitigation. Prior to

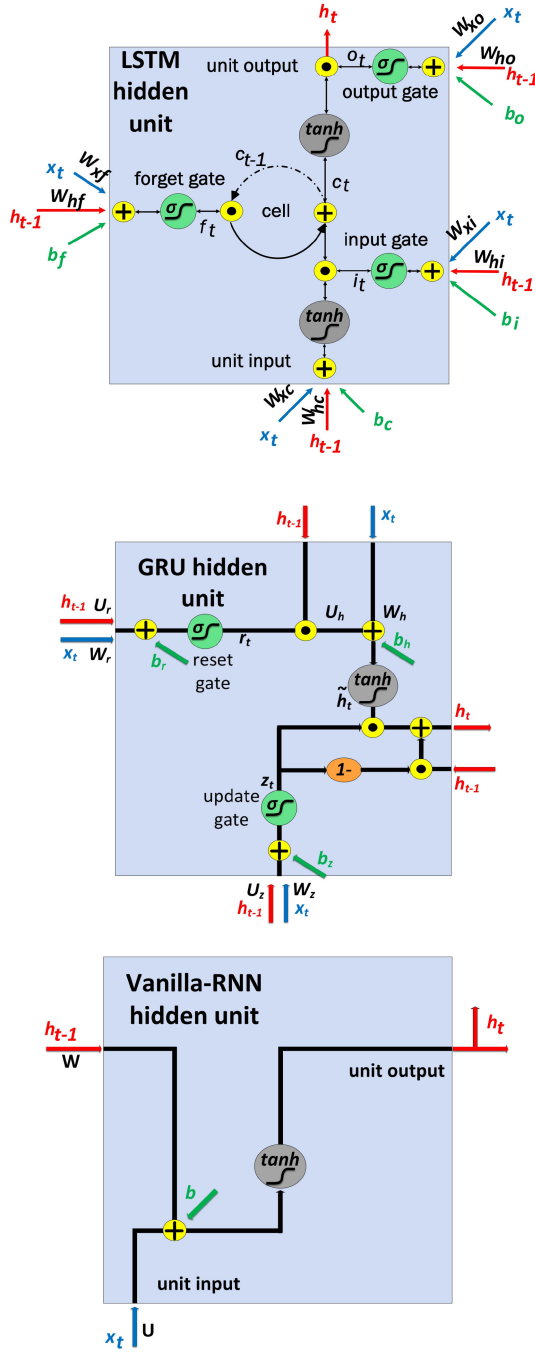


Fig. 2. Conceptual illustration of the LSTM, GRU and Vanilla-RNN units.

any post-processing or demodulation, we first sample the signal and then perform chromatic dispersion compensation with the use of an ideal frequency domain equalizer (FDE). The simulations were conducted with pulse shaping. They incorporate root-raised-cosine (RRC) shaping, with a roll-off factor of 0.2. In this paper we numerically investigate the efficiency of three types of Bidirectional Recurrent Neural Networks (bi-RNN) in compensating fibre communication systems, namely LSTM, GRU and Vanilla-RNN as already reported in the introduction. Fig. 2 illustrates the RNN units that we use whilst Eq. (2)–(4)

indicate how the output h_t is calculated in each case.

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o)$$

$$c_t = f_t * c_{t-1} + i_t * \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c)$$

$$h_t = o_t * \tanh(c_t) \quad (2, \text{LSTM})$$

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z)$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r)$$

$$\tilde{h} = \tanh[W_h x_t + U_h (r_t * h_{t-1}) + b_h]$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h} \quad (3, \text{GRU})$$

$$h_t = \tanh(W h_{t-1} + U x_t + b) \quad (4, \text{Vanilla-RNN})$$

where W matrices contain the weights of connection: f , i and c stands for forget, input, output gate and cell state respectively in the case of (2), W and U matrices contain the weights of connection: z , r , \tilde{h} stands for update, reset gate and candidate activation vector in the case of (3, 4), x_t , h_t , h_{t-1} are input, hidden output, previous hidden output and b are bias vectors. The $*$ operator denotes the element wise product, σ is the logistic sigmoid function and \tanh is the hyperbolic tangent activation function.

The sequential neural model is demonstrated in Fig. 3. The input x_t is the distorted symbol sequence which has the following form $x_{t,L} = [x_{t-k}, \dots, x_{t-1}, x_t, x_{t+1}, \dots, x_{t+k}]$, where L stands for the overall length of the word which is equal to $L = 2k+1$. Thus, for the symbol at time t we also launch k preceding and k succeeding symbols so as to track intersymbol dependencies. The length of L depends on the foreseen channel memory strictly related to accumulated chromatic dispersion. Each symbol in each window contains four values/features (I and Q for both polarizations) as the input Xx-pol and Xy-pol feeding the Bi-RNN layer of L hidden units. In order to calculate bit BER we drive the RNN network output h_t to a Fully Connected Layer of 4 neurons in the case of 16-QAM or 5 neurons in the case of 32-QAM and then $y_{t,L} = [y_{t-k}, \dots, y_{t-1}, y_t, y_{t+1}, \dots, y_{t+k}]$ to a sigmoid layer that carries out the bit-wise estimation $\sigma(y_t)$ for all the symbols at the output. The bit-wise approach [20] is slightly different compared to the symbol-wise approach of [17] and marginally improves the complexity of the model without sacrificing the BER performance.

We train the model using many input and many output symbols, (many to many approach) which predicts simultaneously the same number of symbols y_t as those of input x_t [25]. The RNN models are built, trained and evaluated in Keras with Tensorflow 2.3 GPU backend. In the Keras model, binary cross-entropy is chosen as a loss function and Adam as the optimizer for the BER measurement with the parameters appearing in [26]. We consider 40.000 symbols for training, 20.000 for validation and 60.000 for testing with unknown data. The training stage is executed with batches of 512 words of symbols for optimum balance between memory allocation size and execution time. The maximum forward and backward passes of all training

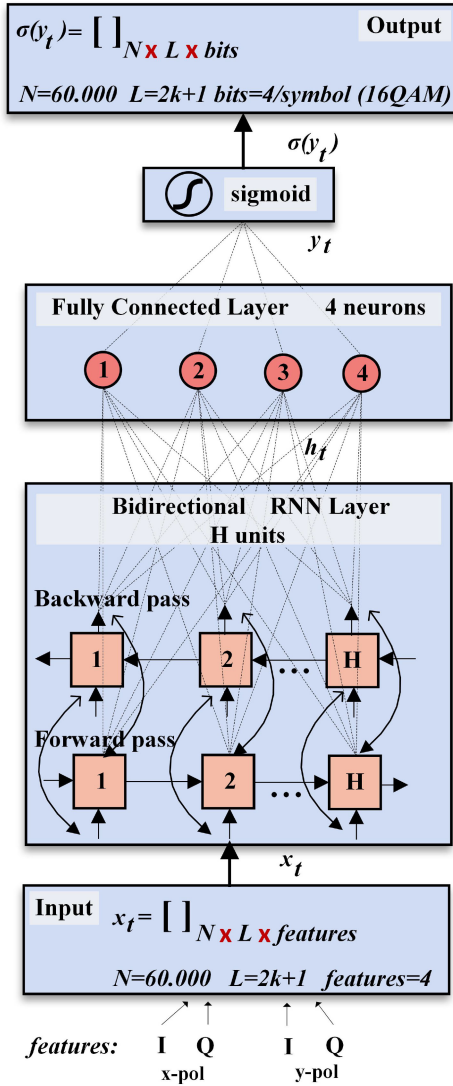


Fig. 3. Bidirectional-RNN model architecture in the case of 16-QAM.

sequences (epochs) are chosen to be 300. To avoid overfitting during training we use “early stopping” when validation loss does not decrease for 20 successive epochs. The RNN processor was applied to the central WDM channel that is in principle the most heavily impaired one and requires only one sample per symbol.

The third-order Volterra nonlinear equalizer used as an alternative to RNN post-processor was simulated as follows: The equalized output $y(n)$ of the $x(n)$ sampled signal from a 3rd order Volterra series can be described according to (5).

$$y(n) = \sum_{i_1=-k_1}^{k_1} w_{i_1} x(n+i_1) + \sum_{i_1=-k_2}^{k_2} \sum_{i_2=i_1}^{k_2} w_{i_1, i_2} x(n+i_1) x(n+i_2)$$

$$+ \sum_{i_1=-k_3}^{k_3} \sum_{i_2=i_1}^{k_3} \sum_{i_3=i_2}^{k_3} w_{i_1, i_2, i_3} x(n+i_1) x(n+i_2) x(n+i_3) \quad (5)$$

where $k_j = (L_j - 1)/2$, in which L_j is the input length of the j -th order. w_{i_1, \dots, i_j} is the weight for each sample of the j -th order [23]. We train the Volterra model trying to minimize error using the least mean square algorithm. We consider 60,000 symbols for training and 60,000 for testing with unknown data. Each input symbol is a vector of four features (I and Q for both polarizations). The following section presents the results of our numerical analysis starting with the comparison between different bi-RNN models.

III. RESULTS AND DISCUSSION

A. Bi-RNN Models - Performance Comparison

First, we compare the three bi-RNN models in terms of BER performance. We test their ability to identify 16-QAM modulated symbol-series of 32 Gbaud, that have been transmitted along 960 km (80 km spans) at 1550 nm. The values of the other propagation parameters are in accordance with Table I. We train the different bi-RNNs with a symbol word of 151 symbols that approximates nonlinear channel memory depending on accumulated dispersion [17]. In Fig. 4, it can be seen that all models perform adequately and equivalently as they improve BER by almost half an order of magnitude ($\text{BER} \sim 1 \times 10^{-3}$) compared to a transmission system that uses only linear equalization. At this point, it must be mentioned that a transmission system that incorporates only linear equalization with the use of FDE, achieves $\text{BER} = 4 \times 10^{-3}$ at minimum for the specific transmission parameters. It can be seen that bi-LSTM offers best performance for a minimum of 14 units, whilst bi-GRU and bi-Vanilla-RNN for 16 units. It is interesting that all models perform equivalently, thus there is no substantial need to use demanding and complex bi-RNN models based on LSTM or GRU units. Bi-Vanilla-RNN topology with almost the same capacity, as dictated by the number of hidden units, but significantly simpler architecture can exhibit similar performance based on findings of Fig. 4. It is well-known that LSTM models use the *cell*, *input*, *output* and *forget gate* structure [27] and GRU models use the *update gate* mechanism, in order to simultaneously prevent the vanishing gradient problem [28] and handle long-term or short-term dependencies [29], namely lags of unknown duration between important events in a time series. The intersymbol interference caused by fibre propagation, as a result of accumulated dispersion, determines channel memory, which remains constant as long as the propagation parameters remain unchanged. Once the network learns the fibre channel memory, it acquires knowledge of the position of the symbols in the sequence and there are no unexpected or time-varying long-term or short-term dependencies as for example exist in rapidly time-varying wireless communication channels. Therefore, the complex aforementioned mechanism of LSTM and GRU becomes unnecessary. Based on the aforementioned finding, the training becomes a process that can be easily undertaken

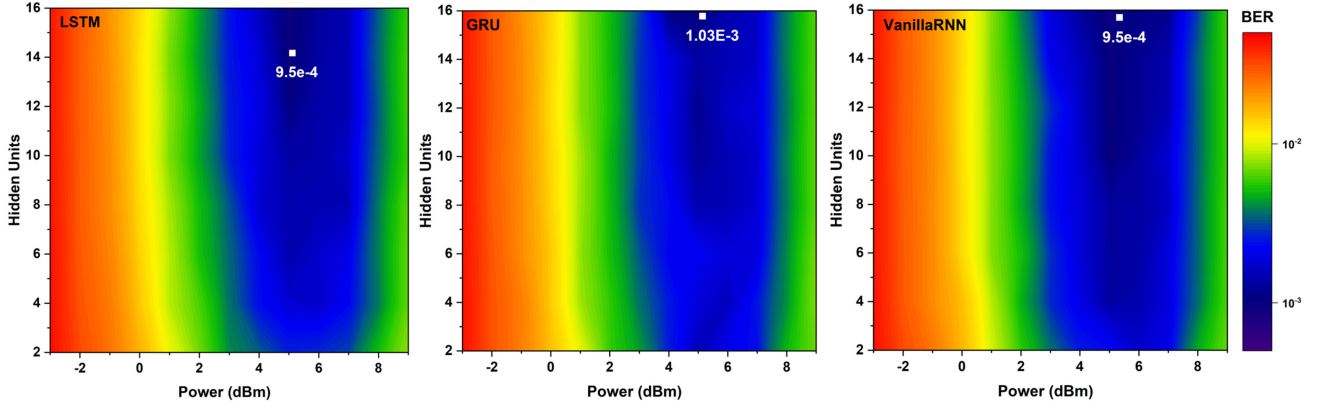


Fig. 4. BER as a function of total launched optical power and number of hidden units for the bi- LSTM (a), bi-GRU (b) and bi-Vanilla-RNN (c) for a 960 km (12 spans of 80 km each) optical transmission link, dispersion of $-21 \text{ ps}^2/\text{km}$ carrying 16-QAM 32 Gbaud signals. The system based only on FDE exhibited minimum BER equal to 4×10^{-3} .

by a bi-Vanilla-RNN model, provided that the symbol word length is long enough to embrace channel memory and nonlinear intersymbol interference. Regarding training efficiency, in all training simulations we carried out, we observed that all RNNs converge more or less to the same loss value but Vanilla-RNN needs about 30% more training epochs than LSTM and GRU, with the latter exhibiting the fastest training. Considering that the optical channel does not require frequent training, the training speed is not critical for the adoption or rejection of a specific RNN model. As already shown in [17], RNN models are robust against power fluctuations or even the modulation format of neighboring channels. Hence, the potential period of re-training relates to the time scales of polarization effects, that is a few ms which corresponds to the duration of hundreds of millions of symbols. Since the system will have been initially trained, re-training will need less symbols (A few thousand) and less time as its purpose will be fine tuning and not training from scratch. This can be accomplished even with the use of a parallel monitoring system which will feed the training unit with a small portion ($< 1\%$) of the real time transmitted and inferred data so as to continuously adjust the weights and keep the system stable.

It must be mentioned that the BER value appearing in Fig. 4 is the one obtained for the central symbol in y_t . Since our RNN models are trained using the many to many approach, one can simultaneously extract all symbols in the word. Fig. 5(a) shows the BER for all symbols in y_t . It becomes evident that the symbols at the center of the word enjoy optimal BER performance. Nevertheless, at least 100 symbols enjoy approximately the minimum BER value ($\sim 1 \times 10^{-3} \pm 0.5 \times 10^{-4}$) also obtained for the central symbol. When we approach the edges of the word, the performance gradually degrades, which confirms the need for training with enough neighboring preceding and succeeding symbols so that non-linear intersymbol interference effects are adequately captured. In the left edge, past symbols are not enough, whilst in the right edge, future symbols are those missing. Fig. 5(b) highlights this behavior in the form of mean BER and its standard deviation when BER is calculated for a portion of the symbols in the word (151 in our case). We formulate these sub-groups of symbols starting from the central symbol

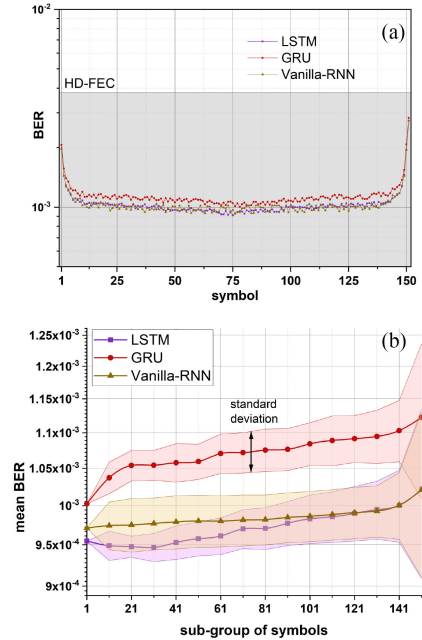


Fig. 5. (a) The value of the BER for each of the 151 symbols of the word, for all RNN models. 100 symbols have approximately the equal minimum BER value, while a total of 100 symbols (green zone) exhibit $\text{BER} < 10^{-3}$. (b) mean BER and its standard deviation calculated in a sub-group of symbols in the word of 151 symbols. Standard deviation explodes when more than 140 symbols are simultaneously detected.

in the word and include the same number of preceding and succeeding symbols. For instance, when sub-group of symbols has a value of 21, then this sub-group contains the central symbol and 10 preceding and succeeding symbols. Based on Fig. 5(b), it becomes obvious that the mean BER remains constant even if 100 out of 150 symbols are simultaneously detected and its standard deviation explodes when subgroups contain more than 140 symbols. Thus, one can easily deduce that the simultaneous detection of 100 symbols in a word of 151 symbols is efficient and will not induce severe degradation of the BER with respect to the case where only the central symbol is decoded.

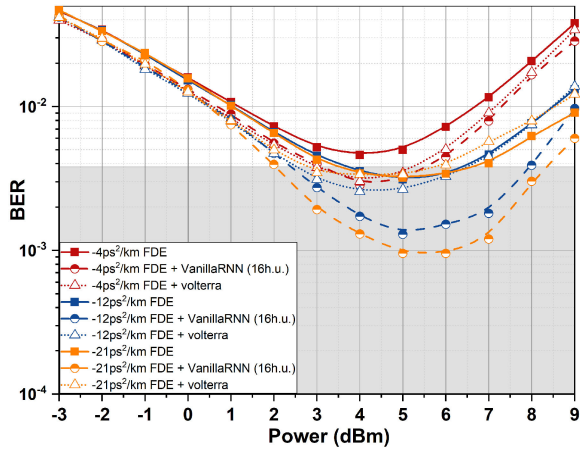


Fig. 6. BER as a function of optical launched optical power for dispersion of -4 , -12 and -21 ps²/km, with linear equalization (FDE) and with a bi-Vanilla-RNN of 16 hidden units or Volterra.

In order to study the efficiency of bi-Vanilla-RNN model for different channel memory scenarios we carried out numerical simulations for -4 , -12 and -21 ps²/km second order dispersion values. Although in real systems a dispersion value different than that of a typical SMF (~ -21 ps²/km) is translated to the use of dispersion compensating modules or other types of fibres that differentiate other critical parameters of the link, here we assume that all the other parameters of the system are not affected in order to identify how the RNN-based equalizer behaves at different channel memories assuming that signal to noise ratio and nonlinearity are kept constant. We have adjusted the word length so as to effectively address memory effects of the highest studied dispersion. With careful training, this does not affect the performance of low memory channels [17]. Apart from FDE equalization, we also conducted numerical simulations using the Volterra nonlinear equalizer. For fair comparison of the two equalizers we use the same memory length ($L = 151$ for RNN and $L_1 = 151$ for the 1st order of Volterra). As far as Volterra is concerned, we also considered $L_2 = 51$, $L_3 = 11$ for the 2nd and 3rd order kernels, two values which are sufficient for equalizing intra-channel nonlinearity impairments and offer the best performance at the least possible complexity as proved by extensive simulations we conducted. In Fig. 6, one can see that FDE compensated transmission systems exhibit BER varying from 4×10^{-3} to 2×10^{-3} with better values appearing in the highest accumulated dispersion values due to the fact that inter-channel nonlinear effects become less intense as dispersion increases. Bi-Vanilla-RNN compensation exhibits significantly better BER at larger accumulated dispersion. Almost identical BER behavior was verified for GRU and LSTM units as well and is related to the coherence time of the channel which is much longer than the symbol period as dispersion increases. Hence inter-channel effects become very slow and easily tracked by the Vanilla-RNN equalizer [17], [30]. Volterra equalizer slightly improves BER compared to linear equalization, however its performance is not improved at larger dispersions. Vanilla-RNN performs much better than Volterra as well as the former seems

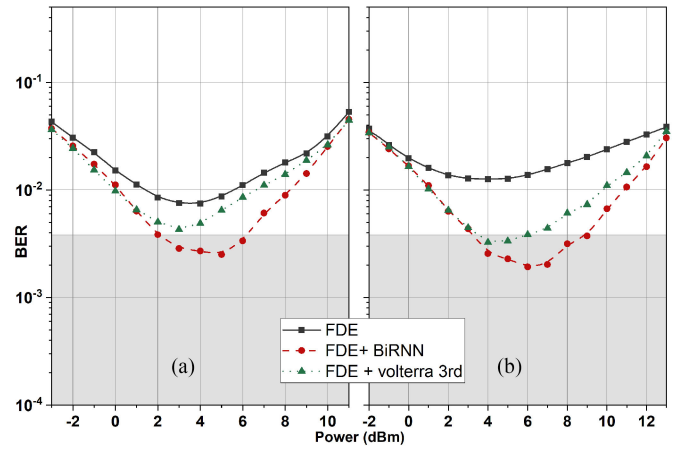


Fig. 7. BER as a function of launched optical power, a) 25 GBaud 32-QAM (960 km, 50 km spacing) (b) 64 GBaud, 16-QAM (640 km, 80 km spacing).

to adequately track inter-channel effects whilst the latter deals with intra-channel effects [31].

B. Operation at Different regimes

We tested the proposed bi-Vanilla-RNN model at different regimes by either increasing the order of modulation or the baudrate. Working with typical single mode fibers in the C-band (1550 nm, -21 ps²/km dispersion) we increased the order of the modulation format to 32-QAM and decreased the baud rate to 25 Gbaud. In Fig. 7a, we observe that only bi-Vanilla-RNN detector is able to perform below the FEC limit, thus showing its systematic superiority in more complex modulation formats as well. It is stated out that GRU and LSTM models in this harsh environment did not exhibit better BER performance than Vanilla-RNN, thus their equivalence was once again proved. These results are not depicted. In order to align our study with the current state-of-the-art transmission systems we increased the baudrate/channel to 64 GBaud with 75 GHz channel spacing, we considered spans of 80 km for a total of 640 km transmission (Fig. 7(b)). The simulation results, even for increased baudrate per channel (~ 500 Gb/s capacity per wavelength) verify that the proposed bi-Vanilla RNN equalizer is capable of mitigating nonlinearities in dispersive channels and once more show its superiority over nonlinear Volterra equalizers.

C. Complexity Analysis

Finally, we investigated the receiver complexity focusing only on bi-RNN and Volterra units. FDE complexity has been analyzed in previous works [17]. In general, the overall bi-RNN equalizer complexity depends on the number of parameters (weights) that each network needs to calculate (see Eq. (2)–(5) and Fig. 1), on the number of hidden units and the length of the input word. According to Fig. 4, LSTM needs at least 14 units whilst GRU and Vanilla-RNN need 16 units to achieve optimal BER, thus the three models are more or less equivalent. Regarding the optimal word length, this does not depend on the selected model architecture, it is strictly related to channel

memory. The only parameter differentiating the three models is the complexity of each unit. Based on (2)–(4), one can easily calculate the number of parameters for each model as follows:

$$bi - RNN_{param} = 2B[H(H + F) + H] + (2H + 1)b \quad (6)$$

where $B = 4, 3, 1$ for LSTM, GRU and Vanilla-RNN respectively indicating the number of gates contained in each RNN unit (see Fig. 2), H the number of hidden units, $F = 4$ the number of input features (see Fig. 3) and b the number of bits ($b = 4$ in the case of 16-QAM). In order to directly compare bi-RNNs with Volterra nonlinear equalizer, it is necessary to calculate the computational complexity in terms of the number of real multiplications. For bi-RNNs we can calculate the number of multiplications, with the use of Eq. (7). We note that the bias is set to zero for all the bi-RNN models, thus it should not be taken into account in the complexity estimation.

$$bi - RNN_{mult} = 2B(FH + H^2)L + 2HbL \quad (7)$$

where L is the length of the input symbol sequence.

The computational complexity $CC_{Volterra}$ of Volterra nonlinear equalizer of eq. (5) in terms of the number of real multiplications is given by Eq. (8) [23].

$$CC_{Volterra}(K, L) = \sum_{k=1}^K \frac{(L_k - 1 + k)!}{(k - 1)!(L_k - 1)!} \quad (8)$$

As Eq. 8 denotes, the number of multiplications explodes as the series order increases [23], this is why we have chosen L_2, L_3 to be smaller than L_1 . In Table IIA, we compare the three models with Volterra in terms of the number of multiplications. The word length is kept at 151 symbols for all equalizers ($L_1 = 151$ for Volterra), whilst for the second and third order we have used $L_2 = 51$ and $L_3 = 11$ as explained in BER performance analysis. It can be seen that among the three bi-RNN models, bi-Vanilla-RNN is the least complex due to a lower B parameter. In the case that RNN predicts/detects only the central symbol, the Volterra is far less complex than all RNN schemes, thus making their use not attractive. Nevertheless, this finding is misleading in the sense that the great advantage of the RNN training method, the “many to many” approach, is not fully exploited here. According to (7), one can easily ascertain that the number of multiplications per symbol in the many-to-many approach reduces with the number of symbols that can be simultaneously detected as shown in Fig. 5. Table II-B and Fig. 8 show the multiplications per symbol needed for the many-to-many approach which was used in this work when multiple symbols are simultaneously decoded. It is evident that taking into account that all RNN schemes offer simultaneous detection of multiple symbols, they are proved to be less complex than Volterra nonlinear equalizer that cannot support a similar mechanism for multiple symbol detection. The bi-Vanilla-RNN turns out to be the least complex when 120 output symbols of 151 are exported at expense of a slight degradation in the BER as depicted in Fig. 5. Hence, the bi-RNN model can be almost 90% lighter than the Volterra nonlinear equalizer. It is worth noting that although there are a lot of studies for complexity reduction for both Volterra equalizers [23] and RNNs, like pruning [32] we decided to identify the complexity of the conventional models. Even if complexity reduction in

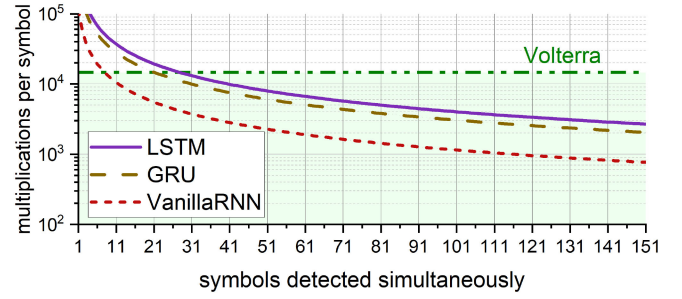


Fig. 8. Number of multiplications per symbol as a function of symbols detected simultaneously.

the order of 70% is taken into account for Volterra based on [23], bi-Vanilla RNN is still preferable in terms of complexity. Pruning is one of the subjects we will study for bi-RNN as well in the near future.

IV. CONCLUSION

In this paper we numerically studied three bi-RNN models, (LSTM, GRU and Vanilla-RNN), as potential fibre nonlinearity compensators in high capacity digital coherent systems. At distances of 1000 km all models exhibited BER improvement of half an order of magnitude compared to systems utilizing exclusively linear equalization and three times lower compared to Volterra nonlinear equalizer. Their efficacy becomes stronger for dispersion unmanaged systems and their superiority over linear and Volterra nonlinear equalization was verified in many transmission scenarios. Among the three models, the bi-Vanilla-RNN exhibits the lowest complexity without lagging behind in BER performance. Finally, we estimated the complexity in terms of the number of multiplications per symbol in the “many to many” training and inference and proved that bi-RNN can be almost 90% less complex than a 3rd order Volterra nonlinear equalizer. Our work explicitly shows that bi-RNNs are a promising post processing method for mid-term deployment in coherent detection transmission systems. In the near future, we intend to study pruning techniques for further minimizing the complexity of RNN models.

REFERENCES

- [1] J. C. Cartledge, F. P. Guimar, F. R. Kschischang, G. Liga, and M. P. Yankov, “Digital signal processing for fiber nonlinearities [invited],” *Opt. Exp.*, vol. 25, no. 3, pp. 1916–1936, 2017.
- [2] G. Li, N. Bai, N. Zhao, and C. Xia, “Space-division multiplexing: The next frontier in optical communication,” *Adv. Opt. Photon.*, vol. 6, no. 4, pp. 413–487, Dec. 2014.
- [3] N. Taengnoi *et al.*, “WDM transmission with in-line amplification at 1.3 μm using a bi-doped fiber amplifier,” *J. Lightw. Technol.*, vol. 37, pp. 1826–1830, 2019.
- [4] I. Phillips *et al.*, “Exceeding the nonlinear-Shannon limit using Raman laser based amplification and optical phase conjugation,” in *Proc. Opt. Fiber Commun. Conf.*, 2014, pp. 1–3.
- [5] A. D. Ellis *et al.*, “Performance limits in optical communications due to fiber nonlinearity,” *Adv. Opt. Photon.*, vol. 9, no. 3, pp. 429–503, 2017.
- [6] S. L. Jansen, D. van den Borne, P. M. Krummrich, S. Spalter, G.-D. Khoe, and H. de Waardt, “Long-haul DWDM transmission systems employing optical phase conjugation,” *IEEE J. Sel. Topics Quantum Electron.*, vol. 12, no. 4, pp. 505–520, Jul./Aug. 2006.

- [7] E. Ip and J. M. Kahn, "Compensation of dispersion and nonlinear impairments using digital backpropagation," *J. Lightw. Technol.*, vol. 26, no. 20, pp. 3416–3425, Oct. 2008.
- [8] Y. Gao, F. Zhang, L. Dou, Z. Chen, and A. Xu, "Intra-channel nonlinearities mitigation in pseudo-linear coherent QPSK transmission systems via nonlinear electrical equalizer," *Opt. Commun.*, vol. 282, no. 12, pp. 2421–2425, Jun. 2009.
- [9] C. Fougstedt, L. Svensson, M. Mazur, M. Karlsson, and P. Larsson-Defors, "ASIC implementation of time-domain digital back propagation for coherent receivers," *IEEE Photon. Technol. Lett.*, vol. 30, no. 13, pp. 1179–1182, Jul. 2018.
- [10] A. Bakhshali *et al.*, "Frequency domain Volterra-based equalization structures for efficient mitigation of intrachannel Kerr nonlinearities," *J. Lightw. Technol.*, vol. 34, no. 8, pp. 1770–1777, Apr. 2016.
- [11] S. K. Turitsyn *et al.*, "Nonlinear fourier transform for optical data processing and transmission: Advances and perspectives," *Optica*, vol. 4, no. 3, pp. 307–322, 2017.
- [12] S. A. Derevyanko, J. E. Prilepsky, and S. K. Turitsyn, "Capacity estimates for optical transmission based on the nonlinear Fourier transform," *Nat. Commun.*, vol. 7, no. 1, 2016, Art. no. 12710.
- [13] E. Giacomidis *et al.*, "Harnessing machine learning for fiber-induced nonlinearity mitigation in long-haul coherent optical OFDM," *Future Internet*, vol. 11, no. 1, pp. 2–11, 2019.
- [14] C. Hager and H. Pfister, "Nonlinear interference mitigation via deep 'neural networks,'" in *Proc. Opt. Fiber Commun. Conf. Expo.*, 2018, pp. 1–3.
- [15] P. Li, L. Yi, L. Xue, and W. Hu, "56 Gbps IM/DD PON based on 10G-class optical devices with 29 dB loss budget enabled by machine learning," in *Proc. Opt. Fiber Commun. Conf.*, 2018, pp. 1–3.
- [16] C. Ye *et al.*, "Recurrent neural network (RNN) based end-to-end nonlinear management for symmetrical 50Gbps NRZ PON with 29dB+ loss budget," in *Proc. Euro. Conf. Opt. Commun.*, 2018, pp. 1–3.
- [17] S. Deligiannidis, A. Bogris, C. Mesaritis, and Y. Kopsinis, "Compensation of fiber nonlinearities in digital coherent systems leveraging long short-term memory neural networks," *J. Lightw. Technol.*, vol. 38, no. 21, pp. 5991–5999, Nov. 2020.
- [18] S. Deligiannidis, C. Mesaritis, and A. Bogris, "Performance and complexity evaluation of recurrent neural network models for fibre nonlinear equalization in digital coherent systems," in *Proc. Eur. Conf. Opt. Commun.*, 2020, pp. 1–4.
- [19] X. Liu, Y. Wang, X. Wang, H. Xu, C. Li, and X. Xin, "Bi-directional gated recurrent unit neural network based nonlinear equalizer for coherent optical communication system," *Opt. Exp.*, vol. 29, pp. 5923–5933, 2021.
- [20] M. Schaedler *et al.*, "Recurrent neural network soft-demapping for nonlinear ISI in 800Gbit/s DWDM coherent optical transmissions," in *Proc. Eur. Conf. Opt. Commun.*, 2020, pp. 1–4.
- [21] M. Schaedler *et al.*, "Soft-demapping for short reach optical communication: A comparison of deep neural networks and volterra series," *J. Lightw. Technol.*, vol. 39, no. 10, pp. 3095–3105, 2021.
- [22] A. Rezaei, J. C. Cartledge, A. Bakhshali, and W.-Y. Chan, "Compensation schemes for transmitter- and receiver-based pattern-dependent distortion," *IEEE Photon. Technol. Lett.*, vol. 28, no. 22, pp. 2641–2641, Nov. 2016.
- [23] N. Diamantopoulos, H. Nishi, W. Kobayashi, K. Takeda, T. Kakitsuka, and S. Matsuo, "On the complexity reduction of the second-order volterra nonlinear equalizer for IM/DD systems," *J. Lightw. Technol.*, vol. 37, pp. 1214–1224, 2019.
- [24] G. Agrawal, *Lightwave Technology: Telecommunication Systems*. New York, NY, USA: Wiley, 2005.
- [25] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.
- [26] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. Conf. Track Proc.*, 2015, pp. 1–15.
- [27] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," in *Proc. 9th Int. Conf. Artif. Neural Netw.*, Sep. 1999, vol. 2, pp. 850–855.
- [28] J. Chung *et al.*, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *Proc. NIPS Workshop Deep Learn.*, Dec. 2014.
- [29] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 855–868, May 2009.
- [30] M. Secondini and E. Forestieri, "On XPM mitigation in WDM fiber-optic systems," *IEEE Photon. Technol. Lett.*, vol. 26, no. 22, pp. 2252–2255, Nov. 15, 2014.
- [31] L. Liu *et al.*, "Intrachannel nonlinearity compensation by inverse volterra series transfer function," *J. Lightw. Technol.*, vol. 30, no. 3, pp. 310–316, Feb. 2012.
- [32] M. Zhu and S. Gupta, "To prune, or not to prune: Exploring the efficacy of pruning for model compression," 2017, *arXiv:1710.01878*.

Stavros Deligiannidis received the B.Sc. degree in physics, the M.Sc. degree in microelectronics, and the VLSI from the National and Kapodistrian University of Athens, Athens, Greece. He is currently working toward the Ph.D. degree with the University of West Attica, Egaleo, Greece, in the field of novel signal processing techniques for optical communication systems under the supervision of Prof. Adonis Bogris. From 2010 to 2019, he was with the Department of Computer Engineering, Technological Educational Institute of Peloponnes, Antikalamos, Greece, where he served as a Lecturer. He has been involved in numerous national and European research projects in the field of optical communications and photonics. His current research interests include optical communications, deep learning, digital signal processing, and parallel computing.

Charis Mesaritis received the B.S. degree in informatics from the Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Athens, Greece, in 2004, the M.Sc. degree in microelectronics from the National and Kapodistrian University of Athens, and the Ph.D. degree in 2011, in the field of quantum dot devices and systems for next generation optical networks from the Photonics Technology and Optical Communication Laboratory, National and Kapodistrian University of Athens. He has actively participated as a Research Engineer or Technical Supervisor in more than ten EU-funded research programs (FP6-FP7-H2020) targeting excellence in the field of broadband communications, cyber-physical security, and photonic integration. He is currently an Associate Professor with the Department of Information and Communication Systems Engineering, University of the Aegean, Mytilene, Greece. He is the author and co-author of more than 60 papers in highly cited peer reviewed international journals and conferences, two international book chapters, whereas he is a regular Reviewer for IEEE, OSA, AIP, and Springer. In 2012, he was awarded a European Scholarship for post-doctoral studies (Marie Curie FP7-PEOPLE IEF) in the joint research facilities of Alcatel-Thales-Lucent in Paris-France, where he worked on intra-satellite communications.

Adonis Bogris (Senior Member, OSA) was born in Athens, Greece. He received the B.S. degree in informatics, the M.Sc. degree in telecommunications, and the Ph.D. degree from the National and Kapodistrian University of Athens, Athens, Greece, in 1997, 1999, and 2005, respectively. His doctoral thesis was on all-optical processing by means of fiber-based devices. He is currently a Professor with the Department of Informatics and Computer Engineering, University of West Attica, Egaleo, Greece. He has authored or coauthored more than 150 articles published in international scientific journals and conference proceedings and he has participated in plethora of EU and national research projects. His current research interests include high-speed all-optical transmission systems and networks, nonlinear effects in optical fibers, all-optical signal processing and all-optical networking, nonlinear effects in lasers and photonic waveguides, mid-infrared photonic devices, neuromorphic computing and cryptography at the physical layer. He is a Reviewer for the journals of the IEEE/OSA.