Research Article | Vol. 30, No. 18 / 29 Aug 2022 / Optics Express 32908
Optics EXPRESS

Attention-aided partial bidirectional RNN-based nonlinear equalizer in coherent optical systems

YIFAN LIU,[1,*] VICTOR SANCHEZ,[1] PEDRO J. FREIRE,[2] JAROSLAW E. PRILEPSKY,[2] MAHYAR J. KOSHKOUEI,[1] AND MATTHEW D. HIGGINS[1]

[1]University of Warwick, Coventry CV4 7AL, UK
[2]Aston Institute of Photonic Technologies, Aston University, Birmingham B4 7ET, UK
[*]yifan.liu.1@warwick.ac.uk

Abstract: We leverage the attention mechanism to investigate and comprehend the contribution of each input symbol of the input sequence and their hidden representations for predicting the received symbol in the bidirectional recurrent neural network (BRNN)-based nonlinear equalizer. In this paper, we propose an attention-aided novel design of a partial BRNN-based nonlinear equalizer, and evaluate with both LSTM and GRU units in a single-channel DP-64QAM 30Gbaud coherent optical communication systems of 20×50 km standard single-mode fiber (SSMF) spans. Our approach maintains the Q-factor performance of the baseline equalizer with a significant complexity reduction of ~56.16% in the number of real multiplications required to equalize per symbol (RMpS). In comparison of the performance under similar complexity, our approach outperforms the baseline by ~0.2dB to ~0.25dB at the optimal transmit power, and ~0.3dB to ~0.45dB towards the more nonlinear region.

1. Introduction

The era of big data stimulates the tremendously increasing demand of data-rich applications such as Internet of Things (IoT) and multimedia services enabled by 5G communications. Service providers and researchers have therefore raised unprecedented interest in achieving reliable high-capacity optical transmissions [1]. However, fiber nonlinearities are still considered to be major impairments that limit the achievable capacity in coherent point-to-point transmission systems [2–4]. Digital signal processing (DSP)-based equalization techniques at the receiver end, such as digital backpropagation (DBP) [5,6], Volterra series nonlinear equalizer (VSNE) [7,8], and perturbation-based nonlinearity equalization [9], to mention a few important example, have been studied extensively over the past few decades to combat optical fiber nonlinear effects. However, these advanced DSP equalization algorithms can only be applied in static connections, as they require prior knowledge of the optical link characteristics (in their original form). Hence, they rely heavily on channel parameters for their computations [10]. Fortunately, machine learning (ML), and especially neural networks (NN), have caught the attention of the research community in the last decade for their ability to learn latent features and underlying connections [11]. They may be used for nonlinear signal equalization without the need to acquire link knowledge in advance. Several NN-based nonlinear equalizers have been proposed as an appended block after chromatic dispersion compensation (CDC) [12–15]. These NN-based nonlinear equalizers adopt several NN structures, such as multilayer perceptrons (MLPs) [12,13], convolutional neural networks (CNNs) [14], and recurrent neural networks (RNNs) [15].

Whilst chromatic dispersion (CD) can be efficiently equalized with linear signal processing, its interplay with nonlinearity results in a nonlinear channel memory effect that requires further

#464159 https://doi.org/10.1364/OE.464159
Journal © 2022 Received 17 May 2022; revised 24 Jun 2022; accepted 19 Jul 2022; published 24 Aug 2022

Optics EXPRESS

equalisation. The use of delay blocks at the input of the NN is first proposed in [12] to take into account channel memory. Such NN architectures consider the received, proceeding, and succeeding symbols together as input and are therefore referred to as "dynamic". To this end, RNNs are popular for their ability to capture temporal dynamic behavior and processes sequential information [16]. A typical unidirectional RNN models the dependence of the current state on the previous state and is principally used for processing data with memory, such as speech and signals affected by inter-symbol interference (ISI) [11]. Moreover, bidirectional RNNs (BRNNs) extend traditional unidirectional RNNs to model the dependence on both past and future states [17]. As a result, several works have investigated RNN-based nonlinear equalizers that consider dynamic sequential input [15,18–21]. These equalizers use long short-term memory (LSTM) [22] and gated recurrent unit (GRU) [23] as the unit structures of the RNNs [15,21] for their capability to learn long term dependencies with gated controls of the input [16]. The notable performance of the BRNN-based equalizers has been validated and confirmed in [19,20].

Despite the great potential of the BRNNs for non-linear equalization, the considerable number of multiplications performed by the gate operations inside the RNN units is still a major concern. The complexity in terms of the number of real multiplications per equalized symbol (RMpS) is strongly correlated with the input sequence length and the number of hidden units of the RNN. Some low-complexity designs have been proposed to reduce such complexity [18,21]. For example, for the bidirectional RNN, GRUs have been used in [21] instead of the more commonly used LSTMs [19]. The work in [18] uses unidirectional RNNs and only includes the preceding half of the symbols as the input. However, these designs either attempt to utilize a new unit structure with a full-length sequence input, or make changes to the input based on trial and error methods rather than understanding the relevance and effects of the input length. It is therefore crucial to investigate and comprehend the role of the input symbols within the input sequence in order to effectively design low-complexity BRNN-based nonlinear equalizers.

To understand the high-level contribution of the preceding and succeeding symbols in the input sequence and their hidden representations to the equalization of the received symbol, we leverage the attention mechanism. Attention is first introduced in neural machine translation (NMT) for sequence-to-sequence tasks [24]. Its original purpose is to search for the most relevant parts of the input sentence for the prediction of target words. This mechanism can be of great use for nonlinear optical signal equalization, where the input symbols can be regarded as an input sequence with memory and the predicted received symbol as the target. [25] integrates attention mechanism for pruning the fully-connected output layer of the nonlinear equalizer for complexity reduction, however there are neither detailed nor clear descriptions of the complexity reduction and the pruning process, and also no rigorous calculations regarding the complexity presented. The attention mechanism can be further explored in such equalizer. In this paper, we consider the Q-factor as the performance metric in a single channel dual-polarization (DP) 64 quadrature amplitude modulation (QAM) coherent optical transmission system transmitting at 30GBaud along 20×50km standard single mode fiber (SSMF) spans. We choose the BiLSTM equalizer in [20] as the baseline model. The main contributions of this paper are summarized as follows:

1) We apply an attention mechanism on both the forward and backward directions of the BiLSTM layer of the equalizer. We leverage the information obtained through the attention block to understand and identify which symbols within the input symbol sequence contribute the most to the equalization.

2) We propose a low-complexity partial BRNN-based equalizer with GRU units, which utilizes the attention information acquired through the attention block. We show that our design has better performance than the baseline structure with less complexity.

The rest of the paper is organized as follows. We briefly review important concepts related to RNNs in Section 2. Section 3. presents the attention-assisted partial BRNN equalizer, where the

Optics EXPRESS

attention mechanism is introduced in Section 3.1, its application in BRNN equalizers in Section 3.2, and the proposed partial BRNN equalizer in Section 3.3. Section 4. shows the numerical system setup and results, including the simulation results with the added attention block and the performance and complexity comparisons between our proposed equalizer and the baseline. Finally, Section 5. concludes the paper.

2. Related work

2.1. Recurrent neural networks

A recurrent neural network (RNN) extends the feed-forward neural network to one that processes arbitrary input sequences by capturing the sequential information i.e. time dependence from the input data. RNN comprises recurrent hidden states, whose activation at each time step is dependent on that of the previous time step [26]. Therefore, RNN can be used to learn, for example, the nonlinear memory of a communication channel and handle inter-channel interference (ISI) [11,16,27]. For the classic, or "vanilla" RNN, each unit takes the hidden state of the preceding unit and outputs a current state [11]. Specifically, a sequence of vectors x_t is taken as the input such that

$$h_t = f(x_t, h_{t-1}), \tag{1}$$

where $h_t \in \mathbb{R}^n$ is a hidden state at time t, and f represents an activation function/nonlinear operation. The hidden states are often transformed into a "context vector" c, which summarizes the hidden states as:

$$c = q(\{h_1, \ldots, h_{T_x}\}), \tag{2}$$

where q is a nonlinear function and T_x is the input length.

It is, however, known that the vanilla RNN can often suffer from the infamous vanishing/exploding gradients problem [11]. To solve this impeding issue, some variations such as LSTMs and GRUs, have been introduced [16], which are briefly reviewed next.

2.2. LSTM and GRU

LSTMs can learn the long-range temporal dependencies of a sequential input [22]. GRUs can also learn such long-term dependencies [24], but with less complexity due to its reduced gate operations, and in our work we will deal with LSTMs and GRUs. The structure details of LSTMs and GRUs and their corresponding gated operations are introduced in [16] and [26]. We consider LSTM as is used in the baseline equalizer, meanwhile GRU has less complicated unit structure, and the performance of it has been validated in [21] for nonlinearity equalization in a 120 Gb/s 64-QAM coherent optical communication system with a transmission distance of 375 km.

2.3. Bidirectional RNN

An RNN can also be used with a reverse input sequence, i.e., from end to start [28]. When used in conjunction with the sequence from start to end, it exploits both preceding and future hidden states. This is done by concatenating the output of two RNNs, which is called the bidirectional RNN (BRNN). Hereinafter, we refer to NN-based equalizers that used bidirectional RNNs as BRNN equalizer, and we differentiate them only when we refer to the specific unit structures used, such as LSTMs and GRUs. We refer to BRNN with LSTMs and GRUs as BiLSTM and BiGRU. A BRNN consists of a forward RNN from the start x_1 to the end x_{T_x} of a sequence and a backward one in reverse order from the end x_{T_x} to the start x_1 [17].

Optics EXPRESS

The hidden state h_j at time j that concatenates its forward hidden state $\vec{h}_j$ and its backward hidden state $\overleftarrow{h}_j$, together with the overall BRNN output hidden states h, are written as:

$$\begin{aligned} h_j &= \left[\vec{h}_j^{\top}; \overleftarrow{h}_j^{\top}\right]^{\top}, \quad \text{where } j = 1, \ldots, T_x, \\ h &= \left[h_1; \cdots; h_{T_x}\right]. \end{aligned} \tag{3}$$

2.4. BRNN-based nonlinear equalization in coherent optical communication systems

In coherent optical communication systems, CD as a linear impairment of optical fibers, could be efficiently negated using linear filtering. However, the interplay between CD and non-linearities along the transmission in optical fibers, which results in nonlinear channel memory, is still uncompensated for [19]. NN-based nonlinear equalizers find hidden patterns and latent features in order to further equalize the received signal after CDC. Such equalizers are usually appended at the coherent receiver side after the DSP process, which performs linear equalization for CDC. The use of delay blocks at the NN input in [12] takes into account the channel memory. As a result, the input signal of such an equalizer is considered to be the desired symbol along with its preceding and subsequent symbols as the input sequence.

BRNN handles well with sequential data, utilizes the hidden states of both the previous and future symbols, and takes advantage of the temporal correlations. Consequently, BRNN-based nonlinear equalizers are proposed to enhance the detection of the received signal with LSTM and GRU units [19–21]. The learning process in BRNN-based nonlinear equalizers can be used for a regression task for the prediction of the received symbols [20], or a classification task, where the targets are categorized in correspondence to QAM constellation points [19,21]. The work in [29] has shown that regression-based learning surpasses classification-based learning in higher Q-factor performance. Therefore, the regression-based learning for the BRNN equalizers is adopted in this paper.

We consider a baseline structure of the BRNN-based nonlinear equalizer in Fig. 1, which is proposed in [20]. It consists of an input symbol sequence, a BRNN layer, a flattening layer, and a dense layer as the output layer. The input to the equalizer is a sequence formed by the received symbol r_i together with its preceding and succeeding k symbols: $\{r_{i-k}, \ldots, r_i, \ldots, r_{i+k}\}$, with a length of $2k + 1$. Each symbol has four features, i.e., the real and imaginary parts, $\{\mathrm{Re}\{r_j\}, \mathrm{Im}\{r_j\}\}$, of both the X and Y polarizations. The RNN layers forward and backward of the BRNN are composed of $2k + 1$ LSTM/GRU units to take into account the i-th input sequence associated with r_i, which we call a full BRNN (F-BRNN). This is because the baseline equalizer takes the full length of the symbol sequence as the input. After the BRNN layer with LSTM units, the output is a sequence of hidden states, $\{h_{i-k}, \ldots, h_i, \ldots, h_{i+k}\}$, with each hidden state concatenating forward and backward hidden states according to Eq. (3). The output sequential hidden states are then flattened by the flattening layer, resulting in a single row vector. The dense layer takes the output of the flattening layer and outputs two values relying on two neurons. The real and imaginary parts $\{\mathrm{Re}\{\tilde{r}_i\}, \mathrm{Im}\{\tilde{r}_i\}\}$ of the final predicted received symbol $\tilde{r}_i$ are the outputs of the equalizer.

2.5. Complexity

In this section, the RMpS of the baseline BRNN-based nonlinear equalizer is introduced based on the calculation in [20] and [21]:

$$C_{unit} = n_{op} n_i + n_{op} n_h + 3, \tag{4}$$

Optics EXPRESS

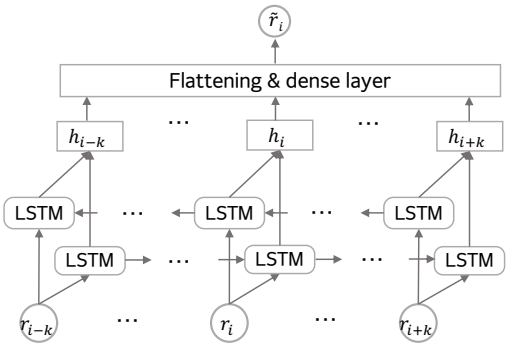


Fig. 1. Baseline BiLSTM-based equalizer

where $n_{op} = 4$ for an LSTM and $n_{op} = 3$ for a GRU as in-unit operations. Therefore:

$$\begin{aligned} C_{unit|LSTM} &= 4n_i + 4n_h + 3, \\ C_{unit|GRU} &= 3n_i + 3n_h + 3, \end{aligned} \tag{5}$$

where n_i is the number of input features, and n_h is the number of hidden units. The complexity for the full BiLSTM and BiGRU C_{full} is given by the following.

$$\begin{aligned} C_{full} &= \underbrace{n_{bi}n_s n_h C_{unit}}_{\text{C}_{\text{input}}} + \underbrace{n_{bi}n_s n_h n_o}_{\text{C}_{\text{output}}} \\ &= 2(2k+1)n_h C_{unit} + 2(2k+1)n_h n_o, \end{aligned} \tag{6}$$

where $n_{bi} = 2$ since we deal with the bidirectional RNNs; n_s is the input symbol sequence length; and n_o is the size of the outputs.

The multiplications can be divided into two terms C_{input} and C_{output} in Eq. (6). It can be seen that the multiplications mostly occur due to the interactions between the input sequence and the gated operations inside the LSTM/GRU units of the BRNN layer. It is therefore pivotal to take good advantage of the recurrent units by reserving only the essential part of the input sequence, in order to design low-complexity nonlinear equalizers. Low complexity RNN equalizer structures have been introduced in [18,21,25,30]. A center-oriented LSTM structure is proposed in [30] that reduces C_{input}, and [25] proposes to use attention mechanism which reduces C_{output}. However, including the baseline structure, there has been no solution that considers the relevance of the preceding and future symbols in the input sequence and their contribution to the output hidden states of the BRNN layer. To this end, we propose a low complexity BRNN-based equalizer based on the attention information gathered from an attention mechanism. We do not only visualize and make sense of the center-oriented LSTM structure in [30], but we also combine the RMpS reduction in both C_{input} and C_{output} as detailed in the next section.

3. Attention-aided partial-BRNN nonlinear equalizer

3.1. Attention mechanism

Attention is a mechanism in NNs which observes a collection of data and selectively focuses on a subset of the collection. It was first applied to sequence-to-sequence learning in [24] and was applied mostly to sequential data to further exploit the importance of each subset among the input data. In other words, attention is one add-on component of a network's architecture, in charge of managing and quantifying the interdependence between the data of interest. General attention

Optics EXPRESS

investigates the interdependence between input and output elements, whilst self-attention deals with finding correlations among input elements [31].

We have a case of general attention to account for the interdependence between the final predicted symbol and both the input symbols and the output hidden states. We simplify the attention mechanism in [24] as our outputs are not sequential. By adding such an attention mechanism, we expect to find the contribution of the input symbols and their hidden representations to the final received symbol prediction. Therefore, we can identify the essential part of the input sequence for training that could lower the computational complexity.

The attention is generally a single- or multi-layer feed-forward NN with trainable weights and biases, which are applied to the output hidden states of the BRNN layer.

In the original attention mechanism [24], an input sequence $\{x_1, \ldots, x_{T_x}\}$ targets an output sequence $\{y_1, \ldots, y_{T_y}\}$. The conditional probability for a certain target output y_i, is defined as:

$$p(y_i|y_1, \ldots y_{i-1}, \mathbf{x}) = g(y_{i-1}, s_i, c_i), \tag{7}$$

where g is a nonlinear, potentially multi-layered, function that outputs the probability of y_i; s_i is a RNN's hidden state for time i computed through: $s_i = f(s_{i-1}, y_{i-1}, c_i)$. Similarly, as defined in Eq. (2), c_i is a vector generated from the sequence of the hidden states for predicting the current target output y_i. c_i is a context vector conditioned for each target y_i; it is computed as a weighted sum of the hidden states $\{h_1, \cdots, h_{T_x}\}$:

$$c_i = \sum_{j=1}^{T_x} \alpha_{i,j} h_j, \tag{8}$$

where the weight $\alpha_{i,j}$ of each h_j is computed by

$$\alpha_{i,j} = \frac{\exp e_{ij}}{\sum_{k=1}^{T_x} \exp e_{ik}}, \tag{9}$$

where $e_{ij} = a(s_{i-1}, h_j)$ is an alignment model which scores how well the inputs around position j and the output at position i match.

We adapt Eqs. (7) to (9) in [24] to our nonlinear equalizer to apply the attention mechanism. Instead of predicting the conditional probability of each target y_i from a sequence of targets, we focus only on the received symbol y_i:

$$y_i = g(c), \tag{10}$$

where

$$c = \alpha * h = [\alpha_1 h_1 \ \cdots \ \alpha_{T_x} h_{T_x}]. \tag{11}$$

The weight α_i of each h_i is calculated by

$$\alpha_i = \frac{\exp\{e_i\}}{\sum_{j=1}^{2k+1} \exp\{e_j\}}, \tag{12}$$

where $e_i = a(h_i)$ is the adapted alignment model and indicates the matching score between the output symbol y_i and the hidden representations $\mathbf{h}$ of the input sequence $\mathbf{x}$. According to [24], we can define the activation function f of the RNN and the alignment model a by choice. A single-layer perceptron (SLP) is selected as our alignment model.

A block diagram of the attention mechanism adopted in this paper with the selected alignment model is shown in Fig. 2. A matrix multiplication is first performed between the hidden input states and a trainable weight matrix $W_a \in \mathbb{R}^{1\times n_h}$ with bias $b_a \in \mathbb{R}^{1\times n_s}$, where n_h is the number of

Optics EXPRESS

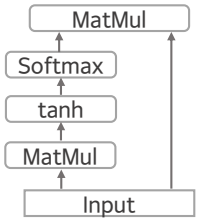


Fig. 2. Block diagram of the adopted attention mechanism.

hidden units, and n_s is the input sequence length, after which a tanh function is applied as the activation function of the SLP:

$$a(h_j) = \tanh(W_a h_j + b_{a_j}). \tag{13}$$

The softmax activation function shown in Eq. (12) is then applied to the alignment model to compute a probability, i.e., the attention score of the hidden states with respect to the final output symbol. The context vector c is then obtained by element-wise matrix multiplication between the attention score α and the hidden states as in Eq. (11). The attention score specifies the amount of attention given to each element of the hidden state sequence that corresponds to that of the input symbol sequence. The block diagram depicting application of the attention block in the BRNN equalizer is shown in Fig. 3(a), with more implementation details shown in Fig. 3(b).

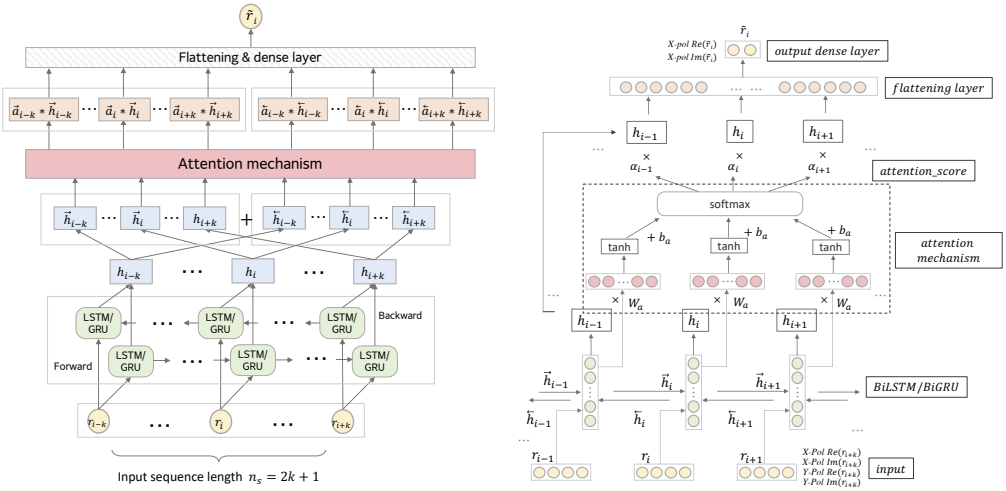


Fig. 3. Block diagram and implementation details of the attention mechanism in BRNN-based equalizer

3.2. *Leveraging attention in BRNN equalizer*

3.2.1. Attention on packed hidden states

As defined in Eq. (3), the conventional way that the hidden states for a F-BRNN are generated is by concatenating the forward and backward or "packed" hidden states for each hidden unit. The block diagram of how the attention mechanism is utilized on this type of F-BRNN is shown in Fig. 3(b). The attention is applied to the summaries of forward and backward hidden states for each time step. In this way, we can see an overall attention that is applied to the hidden

Optics EXPRESS

representations of the input symbols, as the attention block learns how important the hidden states are on each time step in the form of a joint attention score for both directions.

3.2.2. Attention on unpacked hidden states

As the residual channel memory from both preceding and future symbols induced by the interdependence between fiber nonlinearity and chromatic dispersion is often symmetric, we investigate the input sequence from both forward and backward directions individually. This is achieved by concatenating the forward hidden state vector and backward hidden state vector horizontally or "unpacked", as shown in Eq. (14):

$$\begin{gathered}\vec{h} = \left[\vec{h}_{i-k} \ \cdots \ \vec{h}_i \ \cdots \ \vec{h}_{i+k}\right], \qquad \overleftarrow{h} = \left[\overleftarrow{h}_{i-k} \ \cdots \ \overleftarrow{h}_i \ \cdots \ \overleftarrow{h}_{i+k}\right], \\ h = \left[\vec{h}; \overleftarrow{h}\right].\end{gathered} \tag{14}$$

Details of such an implementation in the nonlinear equalizer can be found in Fig. 3(a).

3.2.3. Attention as an auxiliary block

The purpose of applying the attention mechanism is to investigate the relevance between input symbols, their hidden representations, and the outputs. The attention block is leveraged as an auxiliary block to determine which input symbols are selected as the input of the BRNN equalizer and which hidden representations are kept for the final prediction. The attention score is obtained during this auxiliary training process on: (1) the conventional type of concatenated hidden states of the forward and backward RNNs of length $2k+1$ based on Eq. (3) to observe the joint attention on the sequence; (2) the horizontally concatenated forward and backward direction of length $2 \cdot (2k+1)$ based on Eq. (14) to observe the attention in each direction respectively. After getting the attention information, only the relevant part of the input sequence and hidden representations are kept for the proposed BRNN equalizer. The details of such an equalizer are given next.

3.3. Attention-aided partial BRNN equalizer

Despite the fact that attention is directly applied on the hidden representations at the BRNN layer outputs, not only the significance of those hidden representations but also the role of the input symbols in the input symbol sequence can be concluded. As the forward RNN starts from the beginning of the sequence at symbol r_{i-k} to the end, we therefore hypothesize that the attention stops at $r_{i+\vec{a}}$, and that only the first few symbols of the sequence matter. Similarly, as the backward RNN learns the sequential correlation of the sequence in reverse order, the attention also starts from the end r_{i+k} and stops at $r_{i-\overleftarrow{a}}$, as only these symbols matter. In terms of hidden representation, we hypothesize a range $(-\overleftarrow{a}, \vec{a})$ of the input symbol sequence in correspondence to the center received symbol. We obtain an attention score on hidden representations of the BRNN layer, which are generated with the conventional method in Eq. (3). Only a few central hidden representations have a significant score, which verifies our hypothesis. Based on the information we obtain from the attention mechanism, we propose the partial BRNN (P-BRNN) in Fig. 4, where we feed only the relevant symbols in the input sequence to the forward and backward BRNN, respectively. RNNs are applied to a sequence, and the output hidden state of the current time step always relies on that of the previous time step. Therefore, taking the forward RNN as an example, if there is no attention given after a certain element of the sequence, then the input sequence should start from the beginning of the sequence and end at that specific symbol. This is because the time steps after this are redundant and do not contribute to the final prediction. Thus, no unnecessary complexity is introduced, as only the essential part of the input sequence is kept. Moreover, as the attention stops at certain symbols in both forward and backward directions, not all of the output hidden states of the BRNN layer need to be kept and

Optics EXPRESS

fed into the flattening layer. Only $\{h_{i-\overleftarrow{a}}, \ldots h_{i+\vec{a}}\}$ where the attention is given, which are those that correspond to the few adjacent to the center symbols. This means that the corresponding inputs are not necessary for training the network either. The forward and backward attention stops at symbol $r_{i+\vec{a}}$ and $r_{i+\overleftarrow{a}}$ for the received symbol r_i, which indicates that the forward output hidden states from symbol $r_{i+\vec{a}}$ to the end of the sequence, and the backward output hidden states from symbol $r_{i-\vec{a}}$ to the start of the sequence can be trimmed out before the flattening layer. The symbol sequence $\{r_{i-k}, \ldots, r_{i+\vec{a}}\}$ is regarded as the input for the forward RNN, and the symbol sequence $\{r_{i+\overleftarrow{a}}, \ldots, r_{i+k}\}$ as outputs for the backward RNN. As for the input for flattening and dense layer, it is straight forward to keep the hidden state outputs of the centre few symbols, i.e. $\{h_{i-\overleftarrow{a}}, \ldots, h_i, \ldots, h_{i+\vec{a}}\}$ to which the attention is given.

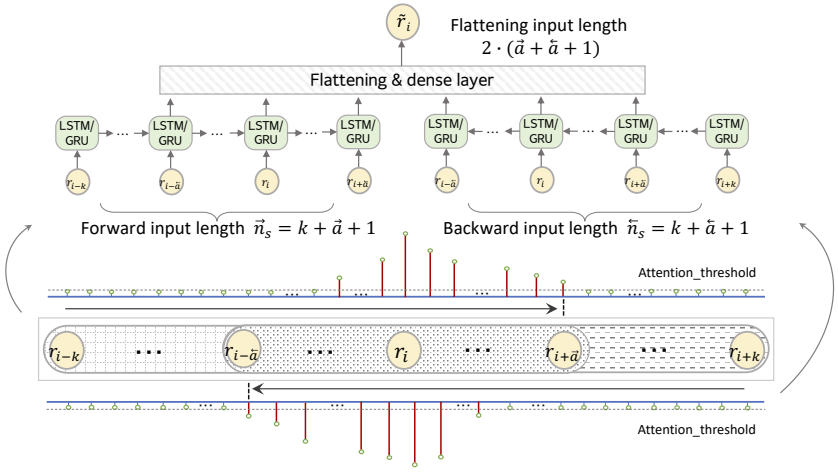


Fig. 4. Proposed partial BRNN-based equalizer which leverages the attention score learned through the attention mechanism during training

3.4. *Complexity reduction*

Similarly to the calculation in Section 2.5, the complexity of our proposed partial BRNN equalizer C_{part} is calculated as follows:

$$\begin{aligned} C_{part} &= (\vec{n}_s + \overleftarrow{n}_s) n_h C_{unit} + 2(\vec{a} + \overleftarrow{a} + 1) n_h n_o \\ &= [(k + \vec{a} + 1) + (k + \overleftarrow{a} + 1)] n_h C_{unit} + 2(\vec{a} + \overleftarrow{a} + 1) n_h n_o, \end{aligned} \tag{15}$$

where $\vec{a}$ and $\overleftarrow{a}$ are the absolute indices of the symbols where the attention of forward and backward RNNs stops, respectively. The implementation details can be found in Fig. 4.

Hence, the total reduction of the complexity in terms of the number of real multiplications per symbol can be calculated based on Eqs. (16) and (15) as:

$$\begin{aligned} C_{reduction} &= C_{full} - C_{part} \\ &= [(n_{bi} n_s C_{unit|LSTM} - (\vec{n}_s + \overleftarrow{n}_s) C_{unit|GRU}] n_h \\ &+ [n_{bi} n_s - 2(\vec{a} + \overleftarrow{a} + 1)] n_h n_o \\ &= [2(2k+1) C_{unit|LSTM} - [2k + 2 + (\vec{a} + \overleftarrow{a})] C_{unit|GRU}] n_h \\ &+ 2(2k - \vec{a} - \overleftarrow{a}) n_h n_o. \end{aligned} \tag{16}$$

Optics EXPRESS

4. Simulation results and discussion

4.1. System setup

We establish our simulations on the single channel transmission of a 30 Gbaud DP-64QAM signal transmitted over 20×50 km SSMF fiber spans with the power swept from −3 dBm to 5 dBm to gain a thorough understanding of how the BRNN-based equalizers perform on a range of transmitted powers.

The received symbols are generated by the simulation that emulates a coherent fiber optic transmission system, shown in Fig. 5. The parameters of the transmission and fiber channel can be found in Table 1. A pseudo random number generator (PRNG) with a Mersenne twister is used to generate 2^{20} random transmitted bits at the transmitter to avoid repetitive patterns of the data. After QAM symbol mapping, data symbols are upsampled, pulse shaped by root-raised cosine (RRC) filters with 0.1 roll-off, modulated by optical in-phase quadrature-phase (IQ) modulator, and sent to the optical fiber channel. At the receiver side, after coherent detection and receiver DSP, i.e., linear filtering that compensates chromatic dispersion, the BRNN-based nonlinear equalizer is applied to further equalize the received signal. The final output predicted received symbol then goes through symbol detection, after which the BER and Q-factor are calculated to evaluate the system performance.

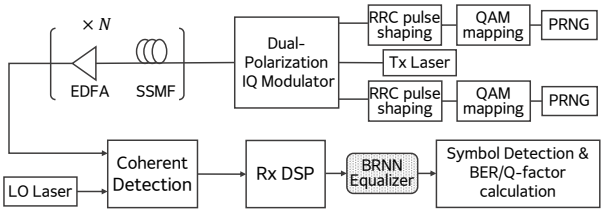


Fig. 5. System setup of our numerical simulations with NN-based nonlinear equalizer for receiver-end processing.

Table 1. Transmission system and fiber parameters

Parameter (Unit)	Value
Modulation format	64QAM
Symbol rate (GBd/s)	30
RRC roll-off factor	0.1
Total fiber length	20 × 50km
Central wavelength (nm)	1550
Fiber attenuation α (dB/km)	0.21
Chromatic dispersion D (ps/nm/km)	16.8
Nonlinear coefficient γ (/W/km)	1.14
Noise Figure (dB)	4.5

4.2. Data pre-processing and model training

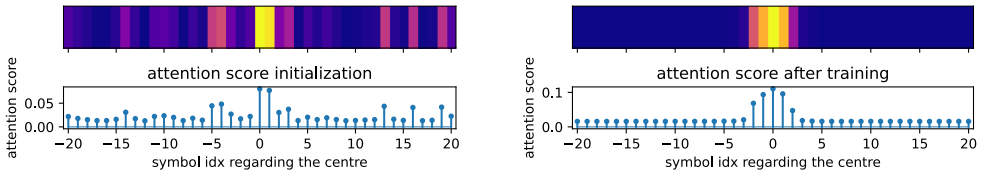
We generate two sets of the transmitted and received symbols from our simulation. In the pre-process stage, the received symbols are separated into groups of 41 symbols with 20 preceding and 20 succeeding symbols adopted in [20]. These are combined into a single full dataset with 1046536 samples. We take the grouped received symbols as the input symbol sequence and the transmitted symbols as the target.

Optics EXPRESS

The training, test and validation datasets comprise 80%, 10% and 10%, respectively, of the total data. The BRNN-based nonlinear equalizers are built, trained, and evaluated using PyTorch 1.9.1. We use mean square error (MSE) as the loss function between the predicted symbols and the received raw symbols, and the Adam optimizer for the gradient descent algorithm in the optimization process. We use the training dataset for training, then validate the model with the validation dataset. As training continues, the Q-factor for the validation dataset is tested with the current trained model. If the computed Q-factor exceeds the last trained model, the best model is updated as the most recently trained model. The test dataset subsequently uses the "best model" to test the Q-factor performance of the BRNN-based equalizer. The maximum training epochs is set to 999, and the training is stopped if the Q-factor for the validation process does not change for a consecutive 300 epochs.

4.3. Experiments of attention on BRNN equalizer

As the hidden representations of both directions of RNNs are utilized for the BiLSTM layer, we therefore apply the joint attention and the individual attention of both forward and backward RNNs in BRNN-based nonlinear equalizers. We first apply the attention mechanism on top of the vertically concatenated LSTM outputs in Fig. 3(b). Figure 6 shows the one-dimensional attention heatmap corresponding to the joint attention score on the bidirectional context features of the BRNN layer computed through the attention block during the training process. Figure 6(a) shows the initialization of the attention score, which is calculated by the softmax function in Eq. (12) after the weight matrix W_a and the bias b_a of the attention layer are initialized, where W_a is drawn from the normal distribution with a standard deviation of 0.001 and b_a is initialized with 0 s.



(a) Attention score initialization.

(b) Final attention score after training.

Fig. 6. Heatmap of joint attention scores on full BiLSTM outputs.

Figure 6(b) shows the final attention score after training, where we can see that generally no attention is given to the symbols on the edges other than the few at the center. The study on the joint attention of both directions gives us good motivation to study individually the attention on each direction of the BRNN.

We experiment with the attention mechanism applied to the hidden states of both the forward LSTM and the backward LSTM layers individually, as shown in Fig. 3(a) and defined by Eq. (14). By feeding $\vec{h}$ and $\overleftarrow{h}$ together into the attention block, we are able to observe and compare the attention score in both directions and the symbol indices of the input sequence where the attention is given. The attention score for forward and backward LSTMs is displayed separately in Fig. 7 after training, where it compares the score value of forward and backward RNNs to check if attention is different for both directions, and compares the symbol indices where the attention stops in the forward and backward directions.

It is shown that attention is equally distributed in both directions. The forward symbol index where attention stops $\vec{a}$ is 3 and the backward symbol index $\overleftarrow{a}$ is also 3. The hidden state of the middle symbol plays the most important role in both directions. Furthermore, the forward and backward attentions manifest similar trends, which validates our findings that it is the nonlinear

Optics EXPRESS

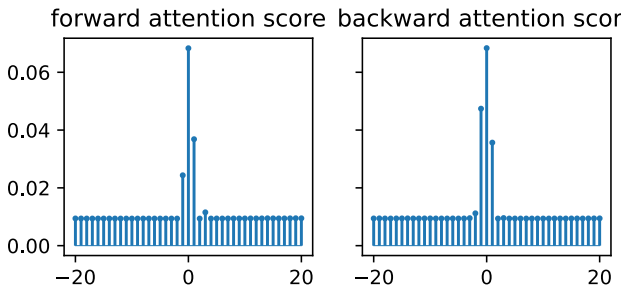


Fig. 7. Attention score comparison between forward and backward RNN using one sequence as an example.

channel memory that occurs at both directions from the center received symbol along the time axis.

4.4. Attention-aided partial BRNN-based nonlinear equalizer

Based on the attention information obtained in Section 4.3, we truncate the input symbols sequence to be fed to the forward and backward RNNs from a full length of $2k+1$ to $k+\vec{a}$ and $k+\overleftarrow{a}$, respectively, as the rest symbols of the sequence do not contribute to the final prediction. The further trimming process occurs before feeding the sequential hidden state outputs of BRNN layer to the flattening layer, where only the hidden states of symbol $r_{i-\overleftarrow{a}}, \ldots, r_i, r_{i+\vec{a}}$ are kept. We use the prefix "trimmed" and suffix "trim" in the rest of the paper to refer to this further trimming process. This is to study the impact on the performance and complexity of this process. As only the necessary symbols of the input sequence are kept as the input of the proposed equalizer, the complexity of the network structure is significantly reduced whilst maintaining the performance. A more detailed complexity comparison can be found in Fig. 10.

4.4.1. Performance after complexity reduction

We first compare the BER performance between the baseline BiLSTM model (F-BiLSTM32) and our proposed model with 32 hidden units (P-BiLSTM32). The models tested are listed in Table 2. Including examining the extra trimming of the center hidden states to be fed into the flattening and dense layers in the proposed partial BRNN equalizers, we compare the results with full BiGRU, partial BiLSTM, and partial BiGRU with extra trim in terms of the hidden representations.

Table 2. Models tested with the same number of hidden units.

Model Type	Input Type/Length	No. hidden units	Trim hidden
F-BiLSTM32	full (2*41)	32	—
F-BiGRU32	full (2*41)	32	—
P-BiLSTM32	partial (2*24)	32	No
P-BiGRU32 trim	partial (2*24)	32	Yes

It is shown in Fig. 8 that our proposed model maintains the performance of the baseline BiLSTM, and a full BiGRU performs slightly worse than the rest. This may be because the symbols that exceed the symbol index where attention stops in both directions not only do not contribute to the sequential hidden state outputs, but also they may instead add to noise. It could also be seen that the BER performance is maintained for the final trimming process.

Optics EXPRESS

Table 3. Models tested under similar complexity

Model Type	Input Type/Length	No. hidden units	Trim hidden
F-BiLSTM32	full (2*41)	32	—
P-BiGRU49 trim	partial (2*24)	49	Yes
F-BiGRU32	full (2*41)	32	—
P-BiLSTM42 trim	partial (2*24)	42	Yes

(a) Complexity based on F-BiLSTM32

Model Type	Input Type/Length	No. hidden units	Trim hidden
P-BiGRU32 trim	partial (2*24)	32	Yes
F-BiLSTM21	full (2*41)	21	—
F-BiGRU24	full (2*41)	24	—

(b) Complexity based on P-BiLSTM32 trim

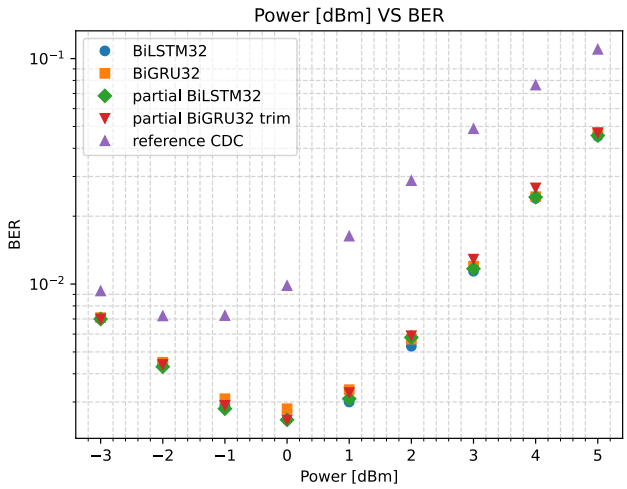


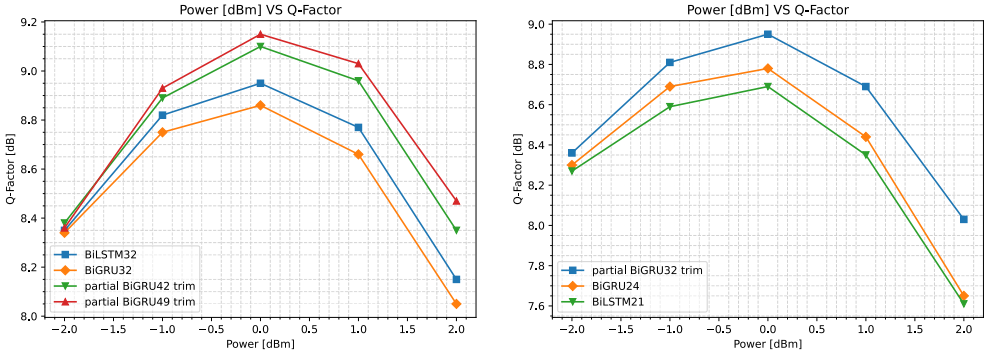
Fig. 8. BER comparison between F-BRNN and P-BRNN

4.4.2. Performance comparison under similar complexity

In terms of further examination of performance enhancement, we conduct our experiment on models with a similar comparable complexity between the baseline F-BiLSTM and our proposed model.

To further demonstrate the benefit from our design, we intentionally chose a slightly lower complexity for our proposed model, as the RMpS is calculated based on several factors and the complexity between two models is hard to precisely match. As shown in Eqs. (6) and (15), the number of hidden units is a constant linear factor in complexity calculation, which is modified in this comparison. We first compare the proposed BRNN equalizer under the similar complexity as the F-BiLSTM with 32 hidden units (referred to as "BiLSTM32"). The models tested are listed in Table 3. We experiment with the following configurations: F-BiGRU with 32 hidden units (BiGRU32), trimmed P-BiGRU with 42 hidden units (partial BiGRU42 trim), and trimmed P-BiGRU with 49 hidden units (partial BiGRU49 trim) in Fig. 9(a). It can be seen from Fig. 9(a) that our proposed model, trimmed P-BiGRU with 49 hidden units, improves the Q factor by 0.2dB at the optimal transmit power 0dBm and 0.3dB in a more highly nonlinear region with transmit power 2.0dBm compared to the baseline BiLSTM.

Optics EXPRESS



(a) Q-factor comparison for F-BiLSTM32

(b) Q-factor comparison based on P-BiGRU32

Fig. 9. Performance comparison under similar complexity between FBRNN and PBRNN.

We then compare the performance of the models with the complexity of the proposed trimmed P-BiGRU with 32 hidden units (partial BiGRU32 trim) as default. We tested F-BiLSTM with 21 hidden units (BiLSTM21) and F-BiGRU with 24 hidden units (BiGRU24). The results of which can be found in Fig. 9(b). It can be seen from Fig. 9(b) that our proposed model trimmed P-BiGRU with 32 hidden units improves the Q-factor by 0.25dB at the optimal transmit power 0dBm and 0.45dB in the more highly nonlinear region with transmit power 2.0dBm compared to the BiLSTM21. The performance enhancement is due to that the larger the number of hidden units, the more features learned in the hidden representations. The Q-factor improvement indicates that performance enhancement is achieved with our proposed design, with even less complexity than other NN structures in comparison. This shows that the complexity of BRNN-based nonlinear equalizer is successfully reduced without decreasing performance.

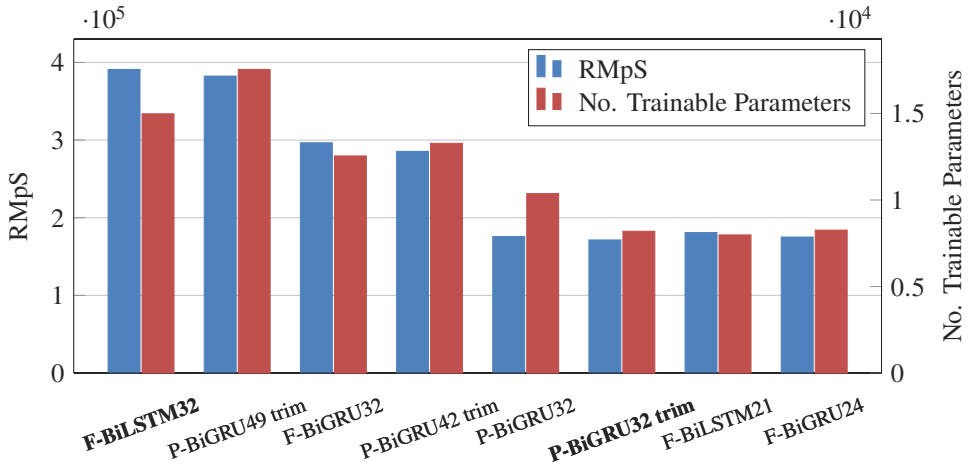


Fig. 10. Complexity comparison for the variety of recurrent NN architectures studied in this paper in terms of RMpS and number of trainable parameters. To demonstrate the potential for complexity simplification when employing our technique, we have highlighted the original BiLSTM with 32 hidden units (F-BiLSTM) and the BiGRU with 32 hidden units following the pruning approach using attention described in this work (P-BiGRU).

Optics EXPRESS

4.4.3. Complexity comparison

A detailed comparison of numerical complexity is conducted in Fig. 10. We calculate the RMpS of the corresponding models based on Eqs. (6) and (15), and we obtain the number of parameters using the summary method inside the torchinfo Python package, which returns the number of trainable parameters in the NN model of interest.

As can be seen, our proposed trimmed partial BiGRU has the lowest complexity with the same performance as the baseline full BiLSTM. The complexity of the BRNN-based nonlinear equalizer is reduced by 56.16% in terms of the RMpS and 45.29% in terms of the trainable parameters. Our proposed approach not only reduces the complexity of the equalizer in terms of RMpS but also reduces the memory storage for the implementation based on the reduced number of trainable parameters. It can also be seen that the final trimming process for hidden states reduces trainable parameters effectively but insignificantly for RMpS.

5. Conclusions

In this paper, we applied an attention mechanism as an auxiliary block in the BRNN-based nonlinear equalizer that equalizes coherent optical signals at the receiver side of a transmission system. We aimed to investigate the significance of the input symbols and their hidden representations in equalizing the received symbol of each input symbol sequence, and proposed a novel design of low-complexity partial-BRNN equalizer, based on the attention score obtained through the attention block. Our results validated symmetric channel memory from both ends of the input sequence, which accumulates along with the transmission and stops at a few symbols beyond the center symbol for both forward and backward directions. Our proposed design was examined in a single-channel DP-64QAM 30Gbaud coherent optical system, in comparison to the baseline BiLSTM equalizer. The proposed BRNN equalizer with GRU units reduced the complexity by 56.16% compared to the baseline. When the equalizers are compared under similar complexity, our approach outperforms the baseline equalizer by ~0.2dB to ~0.25dB at optimal transmit power, and ~0.3dB to ~0.45dB in the more nonlinear region. We conclude that our approach to the attention mechanism provides evidence and an explanation for the importance of symbol-wise nonlinear memory, resulting in an effective evidence-based pruning process of equalizers for optical transmission systems.

Funding. H2020 Marie Skłodowska-Curie Actions (813144); Leverhulme Trust (RP-2018-063); Engineering and Physical Sciences Research Council (EP/N509796/1, EP/R513374/1).

Disclosures. The authors declare no conflicts of interest.

Data Availability. Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

References

1. G. P. Agrawal, *Fiber-Optic Communication Systems* (Wiley, 2021), 5th ed.
2. A. Mecozzi and R.-J. Essiambre, "Nonlinear shannon limit in pseudolinear coherent systems," J. Lightwave Technol. **30**(12), 2011–2024 (2012).
3. E. Agrell, M. Karlsson, A. R. Chraplyvy, D. J. Richardson, P. M. Krummrich, P. Winzer, K. Roberts, J. K. Fischer, S. J. Savory, B. J. Eggleton, M. Secondini, F. R. Kschischang, A. Lord, J. Prat, I. Tomkos, J. E. Bowers, S. Srinivasan, M. Brandt-Pearce, and N. Gisin, "Roadmap of optical communications," J. Opt. **18**(6), 063002 (2016).
4. P. J. Winzer, D. T. Neilson, and A. R. Chraplyvy, "Fiber-optic transmission and networking: the previous 20 and the next 20 years," Opt. Express **26**(18), 24190–24239 (2018).
5. E. Ip and J. M. Kahn, "Compensation of dispersion and nonlinear impairments using digital backpropagation," J. Lightwave Technol. **26**(20), 3416–3425 (2008).
6. A. Napoli, Z. Maalej, V. A. J. M. Sleiffer, M. Kuschnerov, D. Rafique, E. Timmers, B. Spinnler, T. Rahman, L. D. Coelho, and N. Hanik, "Reduced complexity digital back-propagation methods for optical communication systems," J. Lightwave Technol. **32**(7), 1351–1362 (2014).
7. L. Liu, L. Li, Y. Huang, K. Cui, Q. Xiong, F. N. Hauske, C. Xie, and Y. Cai, "Intrachannel nonlinearity compensation by inverse volterra series transfer function," J. Lightwave Technol. **30**(3), 310–316 (2012).

Optics EXPRESS

8. J. Cho and S. T. Le, "Volterra equalization to compensate for transceiver nonlinearity: Performance and pitfalls," in *2022 Optical Fiber Communications Conference and Exhibition (OFC)*, (2022), pp. 1–3.
9. Z. Tao, L. Dou, W. Yan, L. Li, T. Hoshida, and J. C. Rasmussen, "Multiplier-free intrachannel nonlinearity compensating algorithm operating at symbol rate," J. Lightwave Technol. **29**(17), 2570–2576 (2011).
10. E. Yamazaki, A. Sano, T. Kobayashi, E. Yoshida, and Y. Miyamoto, "Mitigation of nonlinearities in optical transmission systems," in *2011 Optical Fiber Communication Conference and Exposition and the National Fiber Optic Engineers Conference*, (IEEE, 2011), pp. 1–3.
11. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, 2016).
12. O. Sidelnikov, A. Redyuk, and S. Sygletos, "Equalization performance and complexity analysis of dynamic deep neural networks in long haul transmission systems," Opt. Express **26**(25), 32765–32776 (2018).
13. M. Schaedler, C. Bluemm, M. Kuschnerov, F. Pittalà, S. Calabrò, and S. Pachnicke, "Deep neural network equalization for optical short reach communication," Appl. Sci. **9**(21), 4675 (2019).
14. D. Wang, M. Zhang, Z. Li, J. Li, M. Fu, Y. Cui, and X. Chen, "Modulation format recognition and osnr estimation using cnn-based deep learning," IEEE Photonics Technol. Lett. **29**(19), 1667–1670 (2017).
15. X. Dai, X. Li, M. Luo, Q. You, and S. Yu, "LSTM networks enabled nonlinear equalization in 50-gb/s pam-4 transmission links," Appl. Opt. **58**(22), 6079–6084 (2019).
16. Z. C. Lipton, J. Berkowitz, and C. Elkan, "A critical review of recurrent neural networks for sequence learning," arXiv preprint arXiv:1506.00019 (2015).
17. M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," IEEE Trans. Signal Process. **45**(11), 2673–2681 (1997).
18. Q. Zhou, C. Yang, A. Liang, X. Zheng, and Z. Chen, "Low computationally complex recurrent neural network for high speed optical fiber transmission," Opt. Commun. **441**, 121–126 (2019).
19. S. Deligiannidis, A. Bogris, C. Mesaritakis, and Y. Kopsinis, "Compensation of fiber nonlinearities in digital coherent systems leveraging long short-term memory neural networks," J. Lightwave Technol. **38**(21), 5991–5999 (2020).
20. P. J. Freire, Y. Osadchuk, B. Spinnler, A. Napoli, W. Schairer, N. Costa, J. E. Prilepsky, and S. K. Turitsyn, "Performance versus complexity study of neural network equalizers in coherent optical systems," J. Lightwave Technol. **39**(19), 6085–6096 (2021).
21. 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. Express **29**(4), 5923–5933 (2021).
22. J. Schmidhuber and S. Hochreiter, "Long short-term memory," Neural Comput. **9**(8), 1735–1780 (1997).
23. K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," arXiv preprint arXiv:1409.1259 (2014).
24. D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv:1409.0473 (2014).
25. A. Shahkarami, M. I. Yousefi, and Y. Jaouen, "Attention-based neural network equalization in fiber-optic communications," in *Asia Communications and Photonics Conference*, (Optical Society of America, 2021), pp. M5H–3.
26. J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," arXiv preprint arXiv:1412.3555 (2014).
27. G. Kechriotis, E. Zervas, and E. S. Manolakos, "Using recurrent neural networks for adaptive communication channel equalization," IEEE Trans. Neural Netw. **5**(2), 267–278 (1994).
28. I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," Advances in Neural Information Processing Systems **27** (2014).
29. P. J. Freire, J. E. Prilepsky, Y. Osadchuk, S. K. Turitsyn, and V. Aref, "Deep neural network-aided soft-demapping in optical coherent systems: Regression versus classification," arXiv preprint arXiv:2109.13843 (2021).
30. H. Ming, X. Chen, X. Fang, L. Zhang, C. Li, and F. Zhang, "Ultralow complexity long short-term memory network for fiber nonlinearity mitigation in coherent optical communication systems," J. Lightwave Technol. **40**(8), 2427–2434 (2022).
31. J. Quinn, J. McEachen, M. Fullan, M. Gardner, and M. Drummy, *Dive into deep learning: Tools for engagement* (Corwin Press, 2019).