applied sciences

MDPI

Article

Optimization Algorithms of Neural Networks for Traditional Time-Domain Equalizer in Optical Communications

Haide Wang [1], Ji Zhou [1,*], Yizhao Wang [1], Jinlong Wei [2], Weiping Liu [1], Changyuan Yu [3] and Zhaohui Li [4,5,*]

[1] Department of Electronic Engineering, College of Information Science and Technology, Jinan University, Guangzhou 510632, China; 1834041007@stu2018.jnu.edu.cn (H.W.); wyz0714@stu2017.jnu.edu.cn (Y.W.); wpl@jnu.edu.cn (W.L.)
[2] Huawei Technologies Duesseldorf GmbH, European Research Center, 80992 Munich, Germany; jinlongwei@gmail.com
[3] Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong, China; changyuan.yu@polyu.edu.hk
[4] State Key Laboratory of Optoelectronic Materials and Technologies, School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510275, China
[5] Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Zhuhai 519082, China
* Correspondence: zhouji@jnu.edu.cn (J.Z.); lzhh88@sysu.edu.cn (Z.L.)

Received: 20 August 2019; Accepted: 15 September 2019; Published: 18 September 2019

Abstract: Neural networks (NNs) have been successfully applied to channel equalization for optical communications. In optical fiber communications, the linear equalizer and the nonlinear equalizer with traditional structures might be more appropriate than NNs for performing real-time digital signal processing, owing to its much lower computational complexity. However, the optimization algorithms of NNs are useful in many optimization problems. In this paper, we propose and evaluate the tap estimation schemes for the equalizer with traditional structures in optical fiber communications using the optimization algorithms commonly used in the NNs. The experimental results show that adaptive moment estimation algorithm and batch gradient descent method perform well in the tap estimation of equalizer. In conclusion, the optimization algorithms of NNs are useful in the tap estimation of equalizer with traditional structures in optical communications.

Keywords: neural networks; optical communications; optimization; equalizer; tap estimation

1. Introduction

In recent years, since deep learning has been applied in image recognition, natural language processing, target tracking, recommendation system, and so on, it is becoming one of the most popular spots for academic research and industrial application [1]. Neural networks (NNs), as an important part of deep learning architectures, can approximate complex nonlinear functions. It turns out that NNs have great potential for solving some intricate problems that cannot be described by analytical methods easily [2]. Research on NNs for optical communication systems has been increasingly popular and has started to be successfully used in many optical communication systems. In coherent optical orthogonal frequency division multiplexing systems, NNs have been tried to mitigate the nonlinear propagation effects [3,4]. Also, NNs have been implemented in intensity modulation and direct detection (IM/DD) systems to overcome both the linear and nonlinear distortions. To reach 50 Gb/s four level pulse amplitude modulations (PAM4) systems over 10 GHz devices, NN was used to optimize the equalization in the receiver side [5]. 32 GBaud and 40 Gbaud PAM8 IM/DD systems with

NN equalizers were demonstrated [6,7]. NN has been considered to be a good solution for eliminating the channel distortion in other communication systems. NN-based equalizers for indoor optical wireless communications [8], 170 Mb/s visible light communications system [9] and modulation format identification in heterogeneous fiber-optic networks were proposed [10]. As a promising type of recurrent neural network, reservoir computing in all-optical implementation enables high-speed signal processing and can set the framework for a new generation of hardware for computing and future optical networks [11].

Although the use of NNs can often bring good results, the high computational complexity of NNs is a problem that cannot be ignored. Generally speaking, the computational complexity of simple multilayer perceptron NN-based equalizers is higher than that of the traditional linear feed-forward equalizer (FFE) and even the Volterra nonlinear equalizer [7,12], not to mention the NNs with more complex structures, such as long short-term memory NN or convolutional NN. The cost of training a NN is very high in terms of computational complexity and size of the training set, which might be not well-suited for some communication systems to perform real-time digital signal processing. Furthermore, there are dangers of overestimating the performance gain when applying NN in systems with pseudo-random bit sequences (PRBS) or with limited memory depths [13]. The use of PRBS may lead to overestimation of the NN performance. However, this issue is beyond the scope of this paper, so it is not discussed here. There is a trade-off between the accuracy and the number of training samples used in the training process for NNs [14]. Since it is believed that more data beats better algorithm [15], scientists in the field of artificial intelligence always use large-scale training data sets to train NNs. As a result, many efficient optimization algorithms have been proposed to ensure fast and stable convergence of minimizing the error function of the NNs models [16–18].

Many problems in many fields of science and engineering can be converted into the optimization problems of maximizing or minimizing objective functions by adjusting the parameters. Gradient-based optimization algorithms are the most commonly used optimization methods in these fields but not the exclusive tools of NN research. The optimization algorithms mainly include the first- and second-order optimization algorithms [19]. However, there are two main limitations in the second-order optimization algorithm, such as Newton's method and its variants. One limitation of the second-order optimization algorithms is that the cost function must be smooth, and the second derivatives are available or numerical approximation is achievable. Another limitation is that the Hessian matrix must be positive definite, and its dimension had better not be too large, taking the computational load into consideration [20]. Thus, the first-order gradient-based optimization algorithms, i.e., gradient descent and its variants are widely used in deep learning problems [21].

If the first partial derivatives of the cost function are available with respect to parameters, the gradient descent method is a very effective optimization method. To properly adjust the parameters, in every training iteration, the optimization algorithm calculates a gradient vector. Then the parameters is changed in the opposite direction of the gradient vector. It is worth noting that the commonly used tap estimation algorithms in time-domain equalizers (TDE) with the traditional structures, i.e., least mean square (LMS) and recursive least square algorithms, are based on first-order gradient of the cost function. Therefore, it is very possible to optimize the TDE with traditional structure by using gradient descent and its variants. In optical communications, especially short-reach optical communications, the distortion models are almost certain. The traditional TDE have been widely applied in optical communication systems and achieved good performance. Therefore, NNs are not always required in some communication systems with deterministic distortion models, taking the cost of training into consideration. In this paper, we propose and evaluate the tap estimation schemes for the equalizer with traditional structures in optical fiber communications using the optimization algorithms commonly used in the NNs. The experimental results show that adaptive moment estimation algorithm and batch gradient descent method perform well in the tap estimation of equalizer. In conclusion, the optimization algorithms of NNs are useful in the tap estimation of equalizer with traditional structures in optical communications.

The rest of this paper is organized as follows. In Section 2, we first review the mathematical model of FFE. Section 3 presents the principle of the proposed tap estimation schemes using optimization algorithms of NNs. The experimental setup is described in Section 4 and the detailed results and discussions are provided in Section 5. Finally, Section 6 concludes the paper.

2. Optimization Problems of Equalizers with Traditional Structures

Since linear transfer function can be generated by the FFE with the feature of easy implementation, it played a very significant role in compensating the channel impairments [22,23]. As depicted in Figure 1, the optimal tap coefficients are estimated by the optimization algorithms. As an application of stochastic gradient descent (SGD) method , LMS algorithm aims to minimize the current square error between the training sample y_t and output of the equalizer s_t, which can be expressed as [24]

$$\underset{\omega_i}{\text{minimize}}(\sum_{i=1}^{N} x_{j-i+1} \times \omega_i - y_t)^2 \tag{1}$$

where N is the number of taps and ω_i is the tap coefficient, for $i = 1, 2, ..., N$. In addition, x_j is the received signal.

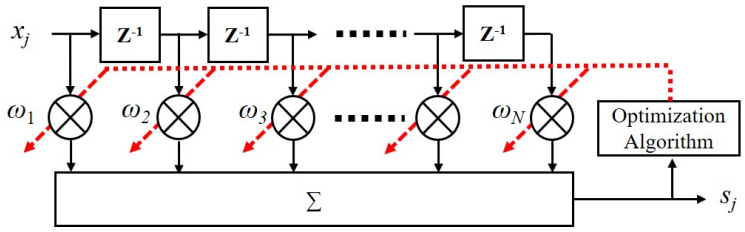

Figure 1. Schematic of an FFE with N taps.

3. Optimization Algorithms for Equalizers with Traditional Structures

3.1. BGD Method

There are three variants of gradient descent, including SGD, batch gradient descent (BGD) and mini-batch gradient descent (Mini-BGD) [21]. The main difference is that they use different number of training sample every training iteration. SGD updates the gradient of error function with respect to the only one training sample, while BGD uses all training samples to perform the parameters updating every iteration. To solve the problem of sharp increase in computation by using all samples every iteration in BGD, partial training samples are used in Mini-BGD every iteration. However, different from deep learning training, which requires large-scale training set, in practical communication systems, it is necessary to use as few training samples as possible. As a result, there is no need to adopt Mini-BGD but using BGD directly with a relatively small batch size. BGD method aims to minimize the mean square error (MSE) of all training samples, which can be expressed as

$$\underset{\omega_i}{\text{minimize}} \frac{1}{M-N+1} \sum_{j=N}^{M} (\sum_{i=1}^{N} x_{j-i+1} \times \omega_i - y_t)^2 \tag{2}$$

where M is the total number of training samples. BGD method using gradients for all training samples to perform just one update of tap coefficients and thus it is a global optimization algorithm. For better convergence, the first $N-1$ training samples are discarded [24]. Matrix form of Equation (2) can be express as

$$\underset{\boldsymbol{\omega}}{\text{minimize}} \frac{1}{M-N+1} (\boldsymbol{R\omega} - \boldsymbol{Y})^T (\boldsymbol{R\omega} - \boldsymbol{Y}) \tag{3}$$

where $\boldsymbol{\omega} = [\ \omega_1 \quad \omega_2 \quad \dots \quad \omega_N\ ]^T$ is the tap coefficient vector of the FFE and $\boldsymbol{Y} = [\ y_N \quad y_{N+1} \quad \dots \quad y_M\ ]^T$ is the desired training vector. The $(M-N+1)$-by-N training matrix $\boldsymbol{R}$ can be expressed as

$$\boldsymbol{R} = \begin{bmatrix} x_N & x_{N-1} & \dots & x_1 \\ x_{N+1} & x_N & \dots & x_2 \\ \vdots & \vdots & \ddots & \vdots \\ x_M & x_{M-1} & \dots & x_{M-N+1} \end{bmatrix} \tag{4}$$

where x_j are the received training samples, for $j = 1, 2, ..., M$. Gradient of MSE with respect with tap coefficient can be calculated as

$$\mathbf{g}(\boldsymbol{\omega}) = \frac{2}{M-N+1}\boldsymbol{R}^T(\boldsymbol{R}\boldsymbol{\omega} - \boldsymbol{Y}). \tag{5}$$

BGD method updates tap coefficient vector $\boldsymbol{\omega}$ in the opposite direction of the gradient, which can be expressed as

$$\boldsymbol{\omega}_t = \boldsymbol{\omega}_{t-1} - \theta \times \mathbf{g}(\boldsymbol{\omega}_{t-1}) \tag{6}$$

where θ is a positive step size and subscript t denotes t-th iteration.

NN researchers have long realized the fact that step size is one of the most difficult hyper-parameters to determine but it is critical to model performance and training costs. Then the following three adaptive step size optimization algorithms are successively proposed for NNs. These algorithms can also be used in tap estimation of equalizers with traditional structures.

3.2. AdaGrad

AdaGrad optimization algorithm scales the step sizes of all tap coefficients inversely proportional to cumulative squared gradient [16]. The cumulative squared gradient can be expressed as

$$\boldsymbol{r}_t = \boldsymbol{r}_{t-1} + \mathbf{g}(\boldsymbol{\omega}_{t-1}) \odot \mathbf{g}(\boldsymbol{\omega}_{t-1}) \tag{7}$$

where the cumulative squared gradient is initialized as N-by-1 zeros vector and $\odot$ denotes the Hadamard product. Gradient can be calculated from Equation (5) and tap coefficients are updated by

$$\boldsymbol{\omega}_t = \boldsymbol{\omega}_{t-1} - \frac{\theta}{\delta + \sqrt{\boldsymbol{r}_{t-1}}} \odot \mathbf{g}(\boldsymbol{\omega}_{t-1}) \tag{8}$$

where a small value δ is used for numerical stability.

3.3. RMSProp

RMSprop is an unpublished, adaptive learning rate optimization algorithm [17]. By changing the cumulative gradient into an exponentially weighted moving average, RMSProp algorithm is derived from AdaGrad algorithm. Compared to AdaGrad, a new hyper-parameter ρ is added to control the scale of the moving average, which can be expressed as

$$\boldsymbol{r}_t = \rho \times \boldsymbol{r}_{t-1} + (1-\rho) \times \mathbf{g}(\boldsymbol{\omega}_{t-1}) \odot \mathbf{g}(\boldsymbol{\omega}_{t-1}). \tag{9}$$

Tap coefficients are also updated as by Equation (8).

3.4. Adam

Adaptive moment estimation (Adam), obtains individual adaptive step sizes for each tap coefficients from estimates of first and second moments of the gradients [18]. The biased first and second moment estimates $\mathbf{m}_t$ and $\mathbf{v}_t$ of $\mathbf{g}(\boldsymbol{\omega}_t)$ are initialized as zeros vector, which are updated as

$$\mathbf{m}_t = \beta_1 \times \mathbf{m}_{t-1} + (1-\beta_1) \times \mathbf{g}(\boldsymbol{\omega}_t), \tag{10}$$

$$\mathbf{v}_t = \beta_2 \times \mathbf{v}_{t-1} + (1-\beta_2) \times \mathbf{g}(\boldsymbol{\omega}_t) \odot \mathbf{g}(\boldsymbol{\omega}_t) \tag{11}$$

where β_1 and β_2 are set to 0.9 and 0.999, respectively. However, they are biased towards zero, especially during the first few steps. Bias-corrected first and second moment estimates counteract biases,

$$\hat{\mathbf{m}}_t = \mathbf{m}_t / \left(1-\beta_1^t\right), \tag{12}$$

$$\hat{\mathbf{v}}_t = \mathbf{v}_t / \left(1-\beta_2^t\right). \tag{13}$$

Tap coefficients are also updated by

$$\boldsymbol{\omega_t} = \boldsymbol{\omega}_{t-1} - \frac{\theta}{\delta + \sqrt{\hat{\mathbf{v}}_t}} \odot \hat{\mathbf{m}}_t. \tag{14}$$

4. Experimental Setups

The performance of the optimization algorithms of NNs that introduced to FFE are investigated by a 129-Gbit/s optical PAM8 system. The Figure 2 shows the experimental setups. At the transmitter, the input bits are first modulated to PAM8 symbols. After added 2000 training samples and 120 synchronized tokens, the digital PAM8 frames are uploaded into a digital-to-analog converter (DAC) with 86-GSa/s sampling rate and 16-GHz 3-dB bandwidth to generate electrical PAM8 frames. There are 82,360 PAM8 symbols per frames and the symbol rate of electrical PAM8 frames is 43 GBaud. A 40-Gbit/s electro-absorption integrated laser modulator (EML) modulates the electrical PAM8 frames to generate the optical PAM8 frames. Next, the generated optical PAM8 signals are launched into 2-km standard single-mode fiber (SSMF). At the receiver, received optical power (ROP) of the signals is adjusted by a variable optical attenuator (VOA). Then the received optical signals are converted into electrical signals by a photodiode (PD). The electrical signals are converted into digital signals by a real-time oscilloscope (RTO) with sampling rate of 80 GSa/s and 3-dB bandwidth of 36 GHz. Finally, off-line processing is implemented to deal with the digital signals, including re-sampling, synchronization, equalization using the FFE with optimization algorithms of NNs, post filter, maximum likelihood sequence detection (MLSD), PAM8 demodulation. The tap number of FFE is set to 101. After equalization, the high-frequency noise is amplified, which greatly degrades the performance of the system. So, a two-tap post filter is adopted to suppress the amplified high-frequency noise [25,26]. Furthermore, a known ISI is introduced by the post filter unavoidably, but it can be eliminated by MLSD based on Viterbi algorithm [27,28].

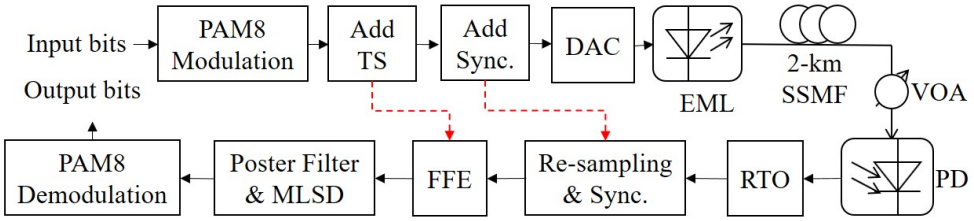

Figure 2. Block diagram of 129-Gbit/s optical PAM8 system. TS, training samples; DAC, digital-to-analog converter; EML, electro-absorption modulator integrated laser; SSMF, standard single-mode fiber; VOA, variable optical attenuator; PD, photodiode; RTO, real-time oscilloscope; Sync, Synchronization.

5. Results and Discussion

In this section, experiment results based on the setup described above are presented and the discussions are also provided. Figure 3 depicts the MSE curves of optimization algorithms versus iteration at ROP of −1 dBm after 2-km SSMF transmission. The iteration numbers of all these optimization algorithms are 120 times. It is clear that after 120 iteration, the MSE curves of using BGD method and Adam algorithm is lower than those of AdaGrad and RMSProp algorithms, which need more iterations to minimize the MSE. The MSE curve of BGD method drops rapidly and steadily because BGD method updates the tap coefficients at the gradient direction. Although the MSE curve of Adam algorithm fluctuates, it drops more quickly. The reason is that it does not update the tap coefficients at the gradient direction but it computes adaptive step size for different tap coefficients from estimates of first and second moments of gradients. It is obvious that tap estimation of FFE using Adam algorithm converges to a lower value of MSE than the other three algorithms. As can be seen from the insets, the diagrams of using BGD method and Adam algorithm are slightly clearer than those of using AdaGrad and RMSProp algorithms, which indicates that BGD method and Adam algorithm may be more effective in tap estimation of FFE.

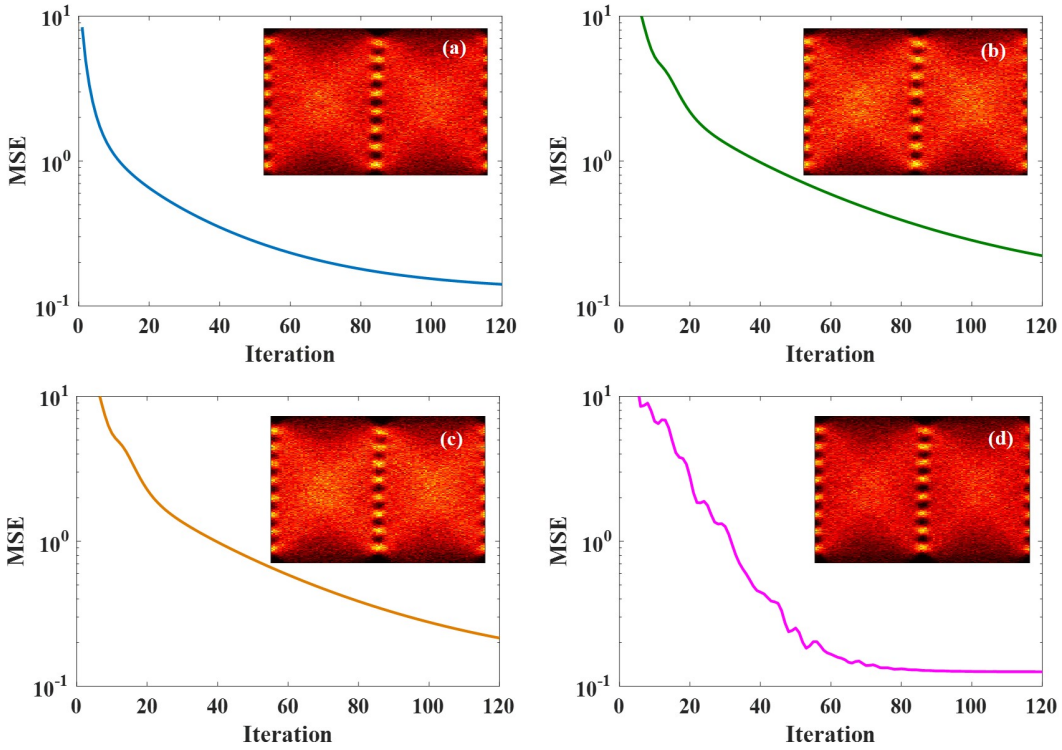

Figure 3. MSE curves of different optimization algorithms applied to 129-Gbit/s optical PAM8 system at ROP of −1 dBm after 2-km SSMF transmission. (**a**) BGD method; (**b**) AdaGrad algorithm; (**c**) RMSProp algorithm; and (**d**) Adam algorithm. Insets are eye diagrams of the equalized signals using the corresponding optimization algorithm.

The bit error rate (BER) performance of 129-Gbit/s PAM8 system versus ROPs at back-to-back (BTB) and 2-km transmission are shown in Figure 4, which can indicate the effectiveness of applying the optimization algorithms of NNs to tap estimation of traditional TDE. As shown in Figure 4a, after BTB transmission, 129-Gbit/s PAM8 system using the above four optimization algorithms have almost the same BER performance. At ROP of −5 dBm, the BER performance is below 7% forward error correction (FEC) limit. Moreover, after 2-km transmission, the BER performance of the system using AdaGrad and RMSProp algorithms are also almost same and below 7% FEC limit when the ROP is greater than or equal to −2 dBm. However, the BER performance of using BGD method is better

than those of using AdaGrad and RMSProp algorithms. Although it is believed that AdaGrad and RMSProp algorithms are effective and practical in deep learning [29], for tap estimation of equalizer in the communication system, it seems that BGD method is better than them in terms of BER performance and computational complexity. The reason may be that these two algorithms are proposed to solve the sparse gradients of the NN and are not particularly superior in the traditional equalizer with simple structure [18]. It is empirically shown that Adam algorithm is usually better than other optimization algorithms in deep learning [18]. In this experiment, Adam also performs better than the other three optimization algorithms. The BER performance of using Adam algorithm is slightly better than that of BGD method and the ROP of using Adam algorithm is more than 1 dB lower than those of using AdaGrad and RMSProp algorithms at the 7% FEC limit. Therefore, the net rate of the system is 117 Gbit/s ($3 \times 43 \times 80{,}240/82{,}360/(1+7\%) \approx 117$ Gbit/s).

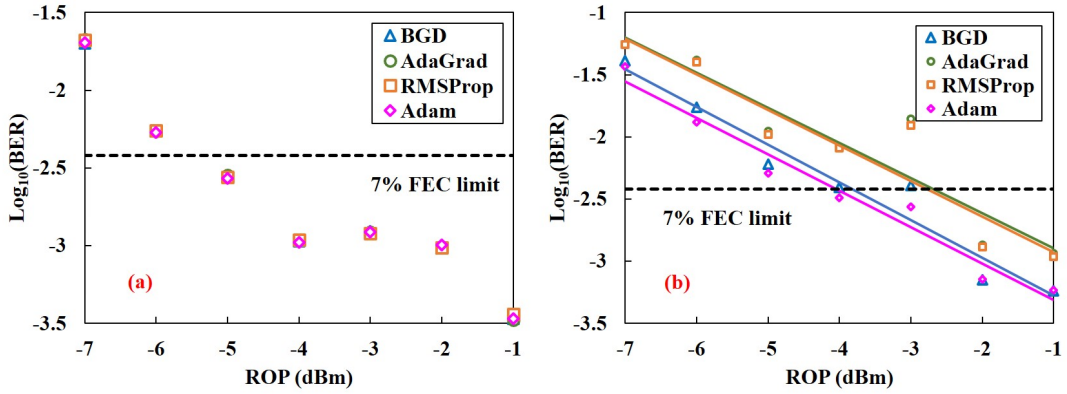

Figure 4. BER performance of 129-Gbit/s PAM8 system versus ROPs at BTB (**a**) and 2-km transmission (**b**) with FFE, post filter, and MLSD. FFE employs BGD (blue triangle), AdaGrad (green rhombus), RMSProp (orange square) and Adam (purple circle), respectively.

Next, the robustness of applying these algorithms to tap estimation of traditional TDE is going to be discussed. Just like LMS algorithm, trial and error is usually required to determine the effective step size of these optimization algorithms, to lead to satisfaction of fast and stable convergence of the tap estimation [30]. As a result, the range of effective step size plays an important role in the robustness of an optimization algorithm. In a general way, the wider the effective step size range of an optimization algorithm is, the better the robustness of the algorithm is. As shown in Figure 5, different optimization algorithms have different effective step sizes and ranges. In this experiment, for BGD method, AdaGrad, RMSProp, and Adam algorithms, the optimal step sizes are about 0.013, 0.3, 0.01 and 0.1, respectively. In general, when the step size is too large, they may not converge or even diverge; but when the step size is too small, they require a lot of iteration [29]. The effective ranges of step size of these four algorithms are respectively about 0.01, 0.87, 0.03 and 7.50. It is obvious that the effective step size range of BGD method is significantly narrower than that of the other three algorithms, because they are adaptive step size algorithms, which improve the robustness of the algorithms at the cost of increasing computational complexity. It is worth noting that Adam algorithm not only has the best BER performance, but also has the widest effective step size range, which means that it is more robust. Their effective ranges of step size in ascending order are as follows, BGD $<$ RMSProp $<$ AdaGrad $\ll$ Adam.

Finally, we analyze the computational complexity per iteration of the above optimization algorithms using in tap estimation. As shown in Table 1, BGD method has the lowest computational complexity, i.e., the minimum number of addition (Add.), multiplication (Mul.) operations and square root (Sqrt.) calculations. Although Adam algorithm has the highest computational complexity compared to the other three algorithms, the difference among them is not obvious. Their complexity in ascending order are as follows, BGD $<$ AdaGrad $<$ RMSProp $<$ Adam.

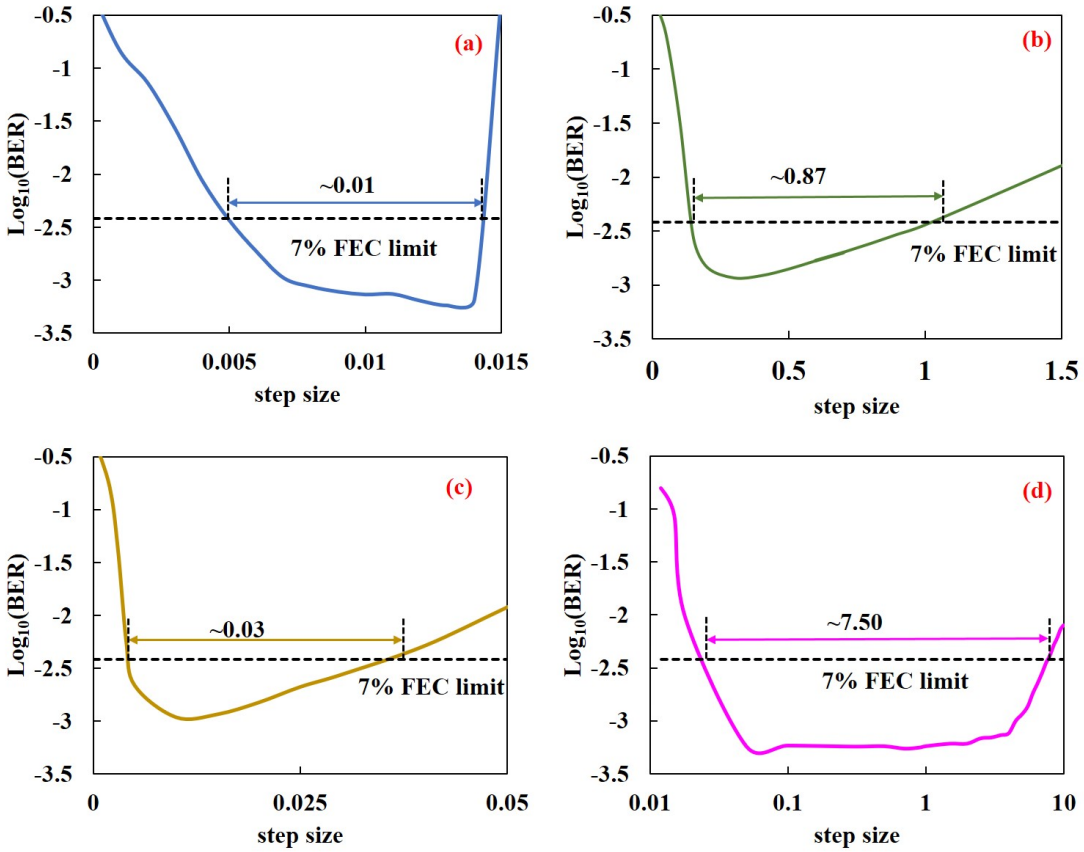

Figure 5. BER versus step size for 129-Gbit/s optical PAM8 system with different optimization algorithms at ROP of −1 dBm after 2-km SSMF transmission. (**a**) BGD method; (**b**) AdaGrad algorithm; (**c**) RMSprop algorithm; and (**d**) Adam algorithm.

Table 1. Computational complexity per iteration of optimization algorithms of NNs introduced to tap estimation of traditional equalizer.

	BGD	AdaGrad	RMSProp	Adam
Add.	$(2N-1)(M-N+1)$	$(2N-1)(M-N+1)+N$	$(2N-1)(M-N+1)+N$	$(2N-1)(M-N+1)+2N$
Mul.	$N[2(M-N+1)+1]$	$N[2(M-N+1)+3]$	$N[2(M-N+1)+5]$	$N[2(M-N+1)+9]$
Sqrt.	0	N	N	N

6. Conclusions

In this paper, we propose and evaluate the tap estimation schemes for the traditional TDE in optical fiber communications using the optimization algorithms commonly used in the NNs. The experimental results show that the optimization algorithms of NNs are also useful in tap estimation of optical communication system. BER performance of 129-Gbit/s PAM8 optical communication system adopting BGD method, AdaGrad, RMSProp, and Adam algorithms are all below 3.8×10^{-3}. It is also shown that Adam algorithm and BGD method perform better in the tap estimation of equalizer. Although Adam algorithm has the highest computational complexity compared to the other three algorithms, its performance is best and it is most robust with the effective step size range of ~7.50. BGD method performs better than AdaGrad and RMSProp algorithms and it is more straightforward to implement, but it is less robust with the effective step size range of ~0.01. In conclusion, the optimization algorithms of NNs are useful in the tap estimation of equalizer with traditional structures in optical communications.

Author Contributions: Conceptualization, H.W., J.Z., J.W., W.L., C.Y. and Z.L.; methodology, H.W., J.Z., J.W., W.L., C.Y. and Z.L.; software, H.W., J.Z. and Y.W.; formal analysis, H.W. and J.Z.; investigation, H.W., J.Z. and Y.W.; writing—original draft preparation, H.W., J.Z. and Y.W.; writing—review and editing, H.W., J.Z., W.L., C.Y. and Z.L.; funding acquisition, J.Z., W.L. and Z.L.

Funding: This research was funded by National Key R&D Program of China (2018YFB1801704); The Science and Technology Planning Project of Guangdong Province (2017B010123005, 2018B010114002); Local Innovation and Research Teams Project of Guangdong Pearl River Talents Program (2017BT01X121); National Science Foundation of China (NSFC) (61525502, 61975242); The Fundamental Research Funds for the Central Universities (21619309).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
2. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef] [PubMed]
3. Jarajreh, M.A.; Giacoumidis, E.; Aldaya, I.; Le, S.T.; Tsokanos, A.; Ghassemlooy, Z.; Doran, N.J. Artificial neural network nonlinear equalizer for coherent optical OFDM. *IEEE Photonics Technol. Lett.* **2014**, *27*, 387–390. [CrossRef]
4. Ahmad, S.T.; Kumar, K.P. Radial basis function neural network nonlinear equalizer for 16-QAM coherent optical OFDM. *IEEE Photonics Technol. Lett.* **2016**, *28*, 2507–2510. [CrossRef]
5. Ye, C.; Zhang, D.; Huang, X.; Feng, H.; Zhang, K. Demonstration of 50Gbps IM/DD PAM4 PON over 10GHz class optics using neural network based nonlinear equalization. In Proceedings of the 2017 European Conference on Optical Communication (ECOC), Gothenburg, Sweden, 17–21 September 2017; pp. 1–3.
6. Gaiarin, S.; Pang, X.; Ozolins, O.; Jones, R.T.; Da Silva, E.P.; Schatz, R.; Westergren, U.; Popov, S.; Jacobsen, G.; Zibar, D. High speed PAM-8 optical interconnects with digital equalization based on neural network. In Proceedings of the 2016 Asia Communications and Photonics Conference (ACP), Wuhan, China, 2–5 November 2016; pp. 1–3.
7. Gou, P.; Yu, J. A nonlinear ANN equalizer with mini-batch gradient descent in 40Gbaud PAM-8 IM/DD system. *Opt. Fiber Technol.* **2018**, *46*, 113–117. [CrossRef]
8. Rajbhandari, S.; Ghassemlooy, Z.; Angelova, M. Effective denoising and adaptive equalization of indoor optical wireless channel with artificial light using the discrete wavelet transform and artificial neural network. *J. Light. Technol.* **2009**, *27*, 4493–4500. [CrossRef]
9. Haigh, P.A.; Ghassemlooy, Z.; Rajbhandari, S.; Papakonstantinou, I.; Popoola, W. Visible light communications: 170 Mb/s using an artificial neural network equalizer in a low bandwidth white light configuration. *J. Light. Technol.* **2014**, *32*, 1807–1813. [CrossRef]
10. Khan, F.N.; Zhou, Y.; Lau, A.P.T.; Lu, C. Modulation format identification in heterogeneous fiber-optic networks using artificial neural networks. *Opt. Express* **2012**, *20*, 12422–12431. [CrossRef]
11. Sorokina, M.; Sergeyev, S.; Turitsyn, S. Fiber echo state network analogue for high-bandwidth dual-quadrature signal processing. *Opt. Express* **2019**, *27*, 2387–2395. [CrossRef]
12. An, S.; Zhu, Q.; Li, J.; Ling, Y.; Su, Y. 112-Gb/s SSB 16-QAM signal transmission over 120-km SMF with direct detection using a MIMO-ANN nonlinear equalizer. *Opt. Express* **2019**, *27*, 12794–12805. [CrossRef]
13. Eriksson, T.A.; Bülow, H.; Leven, A. Applying neural networks in optical communication systems: Possible pitfalls. *IEEE Photonics Technol. Lett.* **2017**, *29*, 2091–2094. [CrossRef]
14. Bottou, L.; Bousquet, O. The tradeoffs of large scale learning. In Proceedings of the Neural Information Processing Systems Conference, Vancouver, BC, Canada, 3–8 December 2007; pp. 161–168.
15. Domingos, P.M. A few useful things to know about machine learning. *Commun. ACM* **2012**, *55*, 78–87. [CrossRef]
16. Duchi, J.; Hazan, E.; Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* **2011**, *12*, 2121–2159.
17. Tieleman, T.; Hinton, G. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks Mach. Learn.* **2012**, *4*, 26–31.
18. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980
19. Battiti, R. First-and second-order methods for learning: between steepest descent and Newton's method. *Neural Comput.* **1992**, *4*, 141–166. [CrossRef]

20. Xiong, X.; De la Torre, F. Supervised descent method and its applications to face alignment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 532–539.
21. Ruder, S. An overview of gradient descent optimization algorithms. *arXiv* **2016**, arXiv:1609.04747.
22. Nielsen, T.; Chandrasekhar, S. OFC 2004 workshop on optical and electronic mitigation of impairments. *J. Light. Technol.* **2005**, *23*, 131–142.
23. Watts, P.M.; Mikhailov, V.; Savory, S.; Bayvel, P.; Glick, M.; Lobel, M.; Christensen, B.; Kirkpatrick, P.; Shang, S.; Killey, R.I. Performance of single-mode fiber links using electronic feed-forward and decision feedback equalizers. *IEEE Photonics Technol. Lett.* **2005**, *17*, 2206–2208. [CrossRef]
24. Haykin, S.S. *Adaptive Filter Theory*; Pearson Education India: Noida, India, 2005.
25. Wang, H.; Zhou, J.; Li, F.; Liu, L.; Yu, C.; Yi, X.; Huang, X.; Liu, W.; Li, Z. Variable-step DD-FTN algorithm for PAM8-based short-reach optical interconnects. In Proceedings of the CLEO: Science and Innovations, San Jose, CA, USA, 5–10 May 2019; paper SW4O.3.
26. Zhou, J.; Qiao, Y.; Huang, X.; Yu, C.; Cheng, Q.; Tang, X.; Guo, M.; Liu, W.; Li, Z. Joint FDE and MLSD Algorithm for 56-Gbit/s Optical FTN-PAM4 System Using 10G-Class Optics. *J. Light. Technol.* **2019**, *37*, 3343–3350. [CrossRef]
27. Li, J.; Tipsuwannakul, E.; Eriksson, T.; Karlsson, M.; Andrekson, P.A. Approaching Nyquist limit in WDM systems by low-complexity receiver-side duobinary shaping. *J. Light. Technol.* **2012**, *30*, 1664–1676. [CrossRef]
28. Zhong, K.; Zhou, X.; Huo, J.; Yu, C.; Lu, C.; Lau, A.P.T. Digital signal processing for short-reach optical communications: A review of current technologies and future trends. *J. Light. Technol.* **2018**, *36*, 377–400. [CrossRef]
29. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: London, UK, 2016.
30. Marquardt, D.W. An algorithm for least-squares estimation of nonlinear parameters. *J. Soc. Ind. Appl. Math.* **1963**, *11*, 431–441. [CrossRef]