photonics

MDPI

Article

Implementation of Pruned Backpropagation Neural Network Based on Photonic Integrated Circuits

Qi Zhang [1,†], Zhuangzhuang Xing [2,†] and Duan Huang [1,*]

[1] School of Computer Science and Engineering, Central South University, Changsha 410083, China; Qi_Zhang@csu.edu.cn
[2] School of Automation, Central South University, Changsha 410083, China; xingz1996@foxmail.com
* Correspondence: duanhuang@csu.edu.cn
† These authors contributed equally to this work.

Abstract: We demonstrate a pruned high-speed and energy-efficient optical backpropagation (BP) neural network. The micro-ring resonator (MRR) banks, as the core of the weight matrix operation, are used for large-scale weighted summation. We find that tuning a pruned MRR weight banks model gives an equivalent performance in training with the model of random initialization. Results show that the overall accuracy of the optical neural network on the MNIST dataset is 93.49% after pruning six-layer MRR weight banks on the condition of low insertion loss. This work is scalable to much more complex networks, such as convolutional neural networks and recurrent neural networks, and provides a potential guide for truly large-scale optical neural networks.

Keywords: optical neural network; network pruning; BP neural network; MRR

Citation: Zhang, Q.; Xing, Z.; Huang, D. Implementation of Pruned Backpropagation Neural Network Based on Photonic Integrated Circuits. *Photonics* **2021**, *8*, 363. https://doi.org/10.3390/photonics8090363

Received: 5 August 2021
Accepted: 26 August 2021
Published: 30 August 2021

1. Introduction

In the past twenty years, deep neural networks have become a milestone in artificial intelligence for its superior performance in various fields, such as autonomous driving [1], stock forecasting [2], intelligent translation [3], and image recognition [4]. However, for the reason of enormous computation in matrix multiplication, traditional central processing units are gradually becoming suboptimal for implementing deep learning algorithms. Silicon photonics [5] provide superior performance in energy consumption [6,7] and computational rate [8] over electronics, which has become an attractive platform. Photonic integrated circuits can easily realize matrix multiplication with the coherence and superposition of linear optics [9]. The advantages of photonics have promoted extensive demonstrations of various high-performance optical functionalities, especially in photon computing [10–15]. The programmable Mach Zehnder interferometers realize optical interference units for optical neural networks [11,16], and MRRs are used for optical matrix-vector multipliers to implement neuromorphic photonic networks [17–19]. In addition, diffractive optics [20–22], space-light modulators [23,24], semiconductor optical amplifiers [25], optical modulators [26,27], and other related optical devices are used to achieve powerful deep learning accelerators, which can perform machine learning tasks with ultra-low energy consumption [28].

The realization of large-scale optical neural networks is still a challenge, though matrix operations can be implemented with special optical components. Excessive integrated optical elements not only increase the manufacturing cost and loss of photonic integrated circuits but also make the adjustment of optical elements more complex [29]. Pruning has been widely used to simplify the model of neural networks [30]. As a typical method of model compression, it can solve the problem of over-parameterization effectively by removing redundant parameters.

In this paper, we demonstrate an optical neural network based on a pruned BP model. MRR banks are used for the matrix computing core can be pruned to the optimum size.

Results indicate that the accuracy of the proposed algorithm is 93.48% in the MNIST dataset with six-layer MRR weight banks pruned. Thus far, we have found that no one has used network pruning to compress optical components. This work can be derived to more complex networks and provides a feasible path toward truly large-scale all-optical neural networks.

2. Materials and Methods

2.1. Backpropagation Neural Network

BP [31] neural network is widely used for strong nonlinear mapping ability in classification. A typical BP neural network includes input layers, hidden layers and output layers. The number of nodes in the input and output layers are always fixed, while the nodes in hidden layers largely affect the performance of neural networks.

The training of a BP neural network includes two processes, forward propagation and backpropagation. Input data are calculated to produce output in forward propagation, which is similar to a typical neural network. The difference comes when expected outputs differ from actual outputs. In backpropagation, deviation is transmitted backward and forward and distributed to all nodes. Parameters of the network are corrected according to the deviation information so that the deviation decreases along the fastest gradient direction. The weights and biases of BP neural networks are updated as follows.

Output of layer L is defined by:

$$y^L = f(x^L) \tag{1}$$

$$x^L = w^L * y^{L-1} + b^L \tag{2}$$

where f is the activation function of neural networks, y^{L-1} represents the output of layer $L-1$, w^L and b^L are weight and bias of layer L.

The evaluation function is defined by:

$$E^N = \sum_{j=1}^{N} (t_j - y_j)^2 \tag{3}$$

where N represents the total number of samples, t_j and y_j represent actual and predictive categories, respectively.

The iterative formulas of weights and biases based on the gradient descent method in the BP model are given as:

$$W_k^L = W_{k-1}^L - \eta \frac{\partial E}{\partial W_{k-1}^L} \tag{4}$$

$$b_k^L = b_{k-1}^L - \eta \frac{\partial E}{\partial b_{k-1}^L} \tag{5}$$

where η is defined as learning rate, a parameter used to control the convergence rate. The value of η is larger at the early stage of training, so weights and biases are updated at a faster speed. Then η is gradually reduced, which is helpful for the convergence of training.

Despite the huge success, a BP neural network needs a lot of computing capacity for its excess parameters. Therefore, network pruning and optical neural networks are introduced to remove redundant parameters and speed up inference.

2.2. Network Pruning

Over-parameterization is widely recognized as one of the basic characteristics of deep neural networks. It ensures that deep neural networks have strong nonlinear mapping ability at the expense of high computational cost and memory occupation. Pruning has been identified as an effective technique to solve this [32]. Weight-elimination [33] is a typical pruning method. It introduces a regularization term to represent structural complexity in network objective function, which can be used to make weights sparser.

Reference [34] introduces the minimum description length (MDL) to describe the complexity of machine learning. According to MDL, the error function is defined by:

$$Err = Err_1 + Err_2 \tag{6}$$

Err_1 has been given in Equation (3).
Err_2 is as follows:

$$Err_2 = \lambda \sum_j \frac{w_j^2 / w_0^2}{1 + \omega_j^2 / w_0^2} \tag{7}$$

where w_0 is called base weight with fixed value, λ is a dynamical parameter that can be adjusted.

Different from the training of the BP model, regularization is introduced into weight adjustment in pruning, which is defined by:

$$w_j(k) = w_j(k-1) + \Delta w_j^1 + \lambda \Delta w_j^2 \tag{8}$$

where

$$\Delta w_j^1 = -\eta \frac{\partial Err_1}{\partial W_j} \tag{9}$$

$$\Delta w_j^2 = -\eta \frac{\partial Err_2}{\partial W_j} = -\frac{2 w_j / w_0^2}{\left(1 + \omega_j^2 / w_0^2\right)^2} \tag{10}$$

Redundant weight decreases continuously until it is small enough to be deleted in training. A hidden node will be deleted when all of its output weight values are close to zero and incorporated into the offset node when all of its input weights are close to zero. Then, we get a simplified model, as shown in Figure 1.

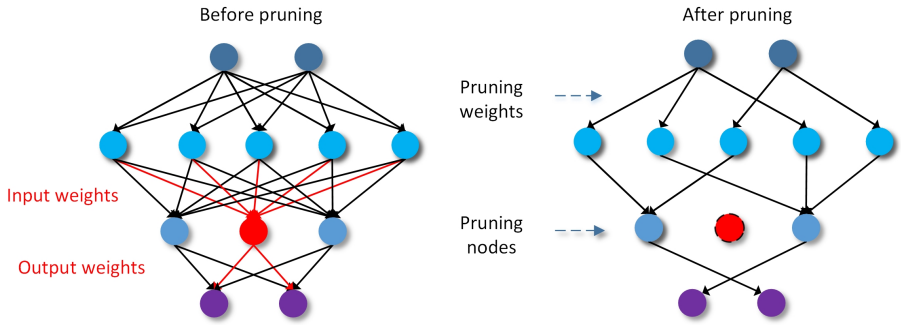


Figure 1. Illustration of pruning: The red node is pruned for all its input weight values that are close to zero with its bias transmitted to the next layer.

2.3. Optical Components and Silicon Photonic Architecture

The MRR banks are the core of the matrix computing. An add-drop MRR consists of two straight waveguides and a circular waveguide, and the straight waveguide at the drop end is set in a curved shape in order to reduce crosstalk sometimes [35]. It has four ports, which are, respectively, called input, through, upload and drop port, as shown in Figure 2a, and z_1 and z_2 serve as coupling regions. Such a silicon waveguide can be fabricated by nanophotonic processing technology, which is compatible with standard complementary metal oxide semiconductor (CMOS) fabrication [36,37]. The add-drop MRR resonance condition is described by:

$$\theta = \beta L = \frac{4\pi^2 n_{eff} r}{\lambda} \tag{11}$$

The parameter θ represents the phase biases through a circular waveguide, and β as the propagation constant of light, which is mainly affected by wavelength λ and the effective index of refraction between the ring and waveguide n_{eff}. The parameters L and r

represent the circumference and radius of the MRR, respectively. Mapping a current in an MRR can change the value of n_{eff}, yielding a shift of the resonance peak [38].

The transfer function of optical intensity going to the drop port from the input is represented by:

$$T_d = \frac{{k_1}^2 {k_2}^2 \alpha}{1 - 2\alpha t_1 t_2 \cos(\varphi) + \alpha^2 t_1^2 t_2^2} \tag{12}$$

The transfer function of the through port light intensity with respect to the input light is

$$T_p = \frac{t_1^2 - 2\alpha t_1 t_2 \cos(\varphi) - \alpha^2 t_2^2}{1 - 2\alpha t_1 t_2 \cos(\varphi) + \alpha^2 t_1^2 t_2^2} \tag{13}$$

where parameters t_1 and t_2 represent the transmission coefficients of z_1 and z_2, k_1 and k_2 represent the mutual coupling factors, respectively. The parameter α defines the loss coefficient in the ring waveguide. Figure 2c shows an MRR without resonance, and the light is all of the output to the through port. Figure 2d shows an MRR in the resonant state, and the light from the straight waveguide is transmitted to the ring. The effective refractive index between the waveguide and the MRR circle causes the phase shift of the light, which interferes with the intensity of the original light and finally outputs to the drop port.

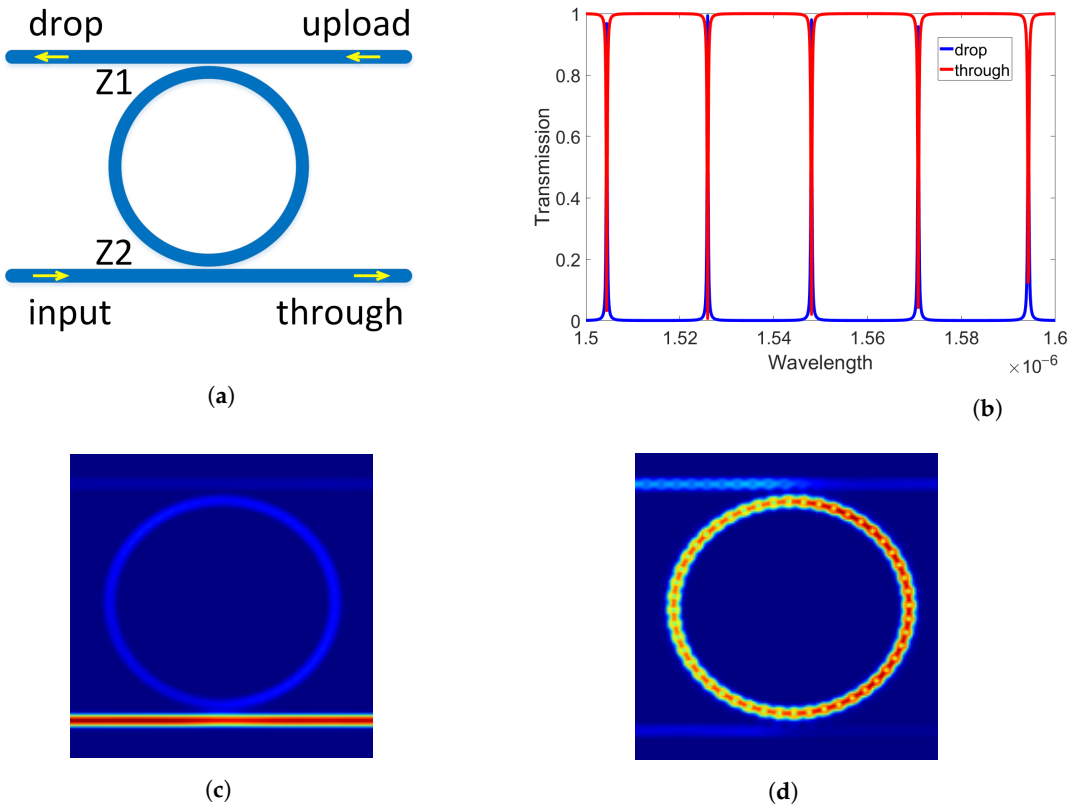


Figure 2. (**a**) An add-drop MRR, which includes two coupling regions and four ports. (**b**) The transfer function of T_d-T_p. The light has five resonances in different wavelengths, and the red curve represents the through port, and the blue one is the drop port. (**c**) The state of add-drop MRR without resonance, light is transmitted from the input to the through port. (**d**) The state of resonance. The light from the straight waveguide is transferred into the ring, and the effective index of refraction between the waveguide and the MRR and the circumference of the MRR cause the light to have a phase shift, thereby the intensity of the original light is interfered and finally outputs from the drop port.

Assuming that the input light has an amplitude of E_0, and the coupling losses are negligible, we can derive the following formulas for the light intensity of the drop and through ports:

$$I_d = (T_d)|E_0|^2 \tag{14}$$

$$I_p = (T_p)|E_0|^2 \tag{15}$$

Figure 2b gives the transfer function of MRR with a radius of 5 µm , and the parameters of the coupling region are identical ($k_1 = k_2 = 0.18$). Different wavelengths of light have different resonance conditions, and the red curve represents the through port, and the blue one is the drop port.

In this work, we demonstrate the matrix operation of BP with MRR banks. Figure 3 illustrates an optical backpropagation network architecture. Suppose we handle two-dimensional matrices $D \times R$, we need R lasers with generating M wavelengths, where M is the type of pixel. We use R modulators to represent the value of each pixel, each of which keeps the light intensity of the corresponding carrier proportional to the serialized input pixel value [27]. Then, WDM will multiplex the R lasers and split them into D separate lines. There are D MRR-arrays, where each array has R MRRs on every line, and we can get a $D \times R$ weight matrix. Thus, the standard form of multiplication and accumulation in BP is represented by:

$$y_d = \sum_{i=1}^{k} A_i F_i + \sum_{i=1}^{k} E_0 F_i \tag{16}$$

where y is the output of layer d, A_i is the light intensity of line i after modulating and multiplexing, F_i, as a particular weight according to MRR weight banks and balanced photodiodes, can be described by:

$$F = g(T_d - T_p) \tag{17}$$

where g is the gain of an amplifier (TIA) to ensure that $T_d - T_p$ is not limited in the range -1 to $+1$ [19]. The sum of $E_0 F_i$ is a predictable bias.

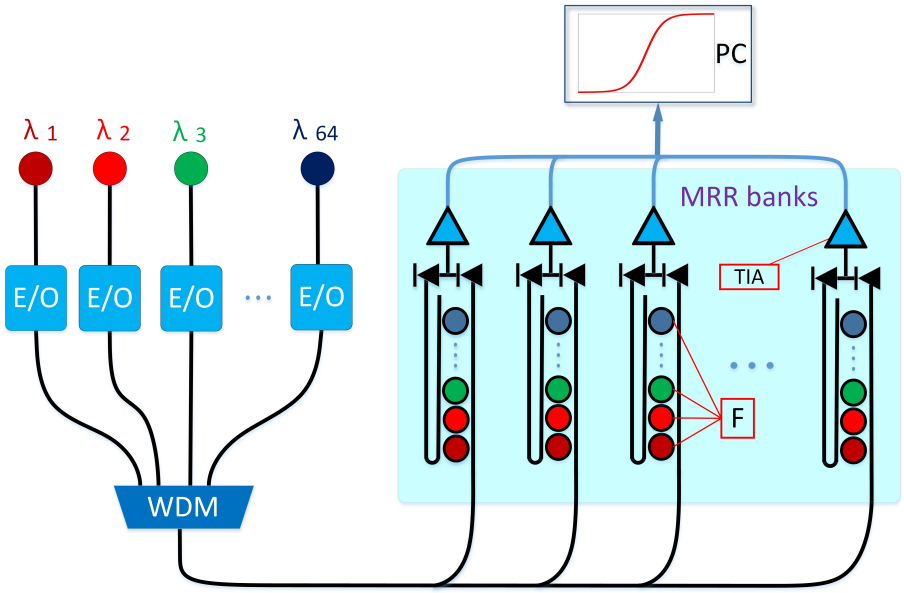


Figure 3. The optical backpropagation network architecture. λ represents different wavelengths. EO represents electro-optic modulators. WDM, wavelength division multiplexing. MRR banks consist of multiple add-drop MRRs in parallel. PD, detector. TIA, amplification. PC, digital computer.

2.4. Photonic Pruning Neural Network

In this work, we trained a neural network based on a pruned BP model to perform image recognition on the MNIST dataset, and Figure 4 depicts this model in detail.

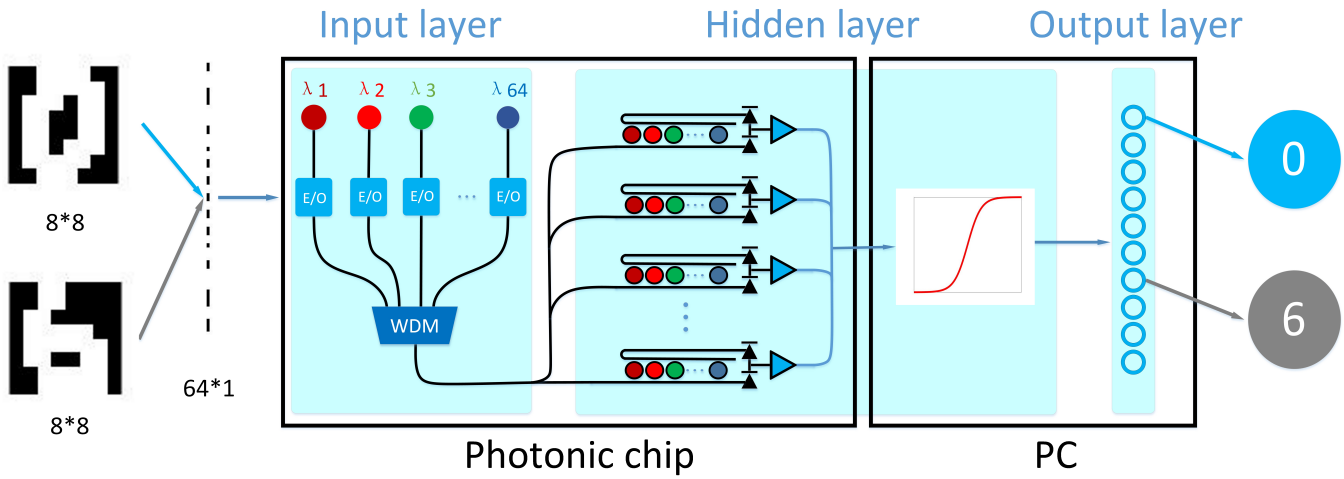
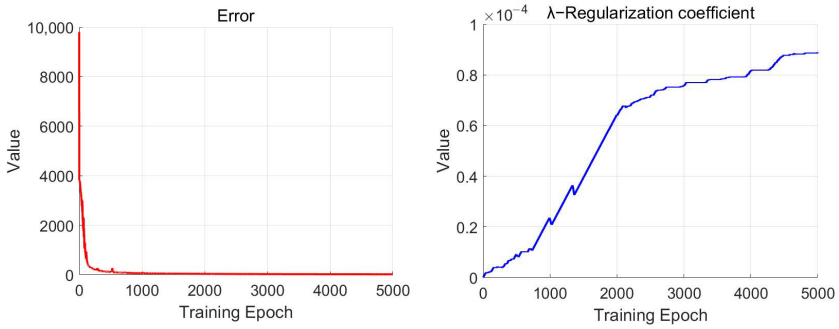


Figure 4. Diagram of the three-ply optical neural network solving MNIST. The two-dimensional images of 8 × 8 are converted into a one-dimensional image of 64 × 1, and they are put into the photoelectric modulators and modulated with 64 optical carrier signals.

The pruned BP model parameters are pretrained in a digital computer (PC). An optical neural network is used to perform matrix multiplication in inference. First, we prune the model during training based on weight-elimination, and weights with biases of the pruned model are uploaded to the optical neural network to calculate optical device parameters. The two-dimensional images of 8 × 8 in the MNIST dataset are converted into a one-dimensional image of 64 × 1, which are put into the photoelectric modulators and modulated with 64 optical carrier signals. When 64 multiplexed optical signals are transmitted through 50 wave wires to an array with 64 MRRs, the signal from the weight banks is output to the digital computer for nonlinear activation. Finally, the transformed vector is fed to the full connection layer with 10 nodes, where the result of the MNIST classification appears.

3. Results and Discussion

In the pruning experiments, we count the error and regularization coefficient of each training during pruning, as shown in Figure 5a,b. It is obvious that the error curve in the training process is relatively smooth, while the regularization coefficient is not, which means the process of weight differentiation does not lead to sharp fluctuations in training and has little effect on pruning results. However, we find that the pruning of small-scale neural networks is not always stable. The data of the experiment shows that the accuracy of the pruned model can still reach 85.25% with half of the original information retained. A small-scale neural network model is easy to fall into local extremum, which means that it is difficult to guarantee its global convergence in training. Although small-scale model compression for neural networks is feasible, the compression performance is not obvious compared with large-scale deep neural networks for the lack of redundant parameters. Figure 6 depicts the prediction accuracy of the BP network through pruning different nodes, and we can find, when 6 neurons are trimmed in the hidden layer with 50, the prediction accuracy is the best, reaching 95.39%.



(**a**) (**b**)

Figure 5. (**a**) Variation of error curve during pruning. (**b**) Variation of regularity coefficients during pruning.

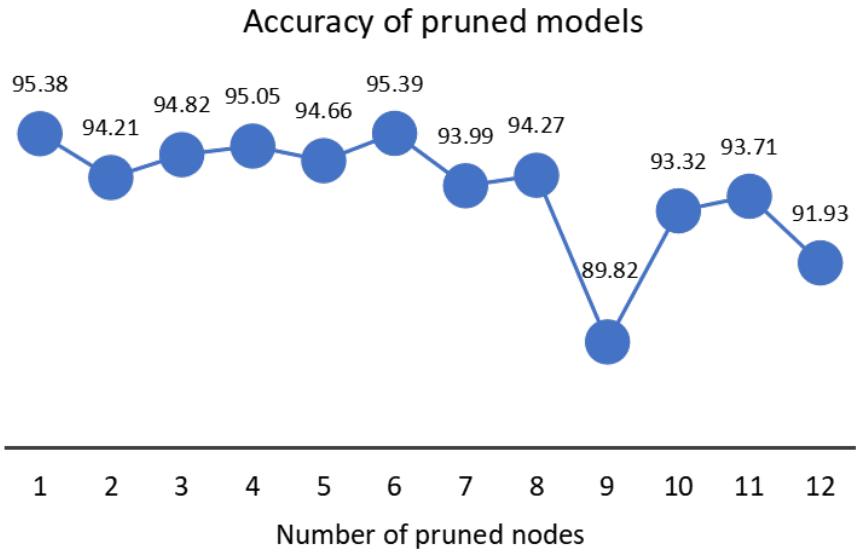


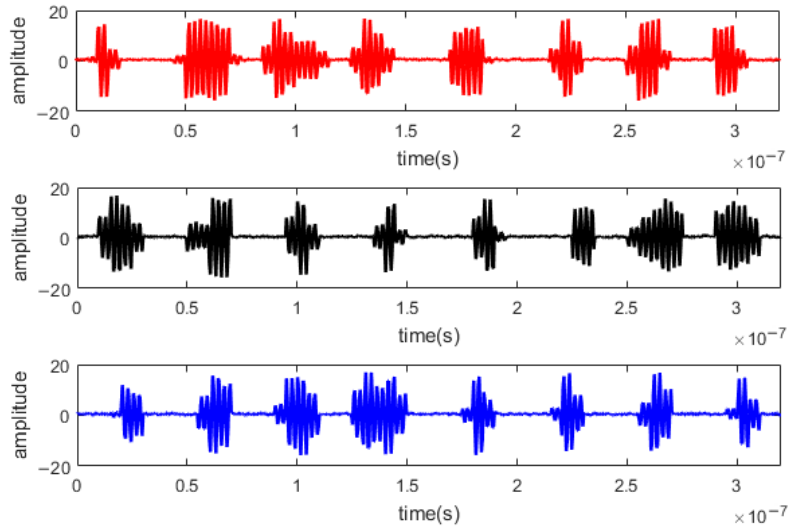
Figure 6. Prediction accuracy of BP network through pruning different nodes. The horizontal ordinate represents the layers of pruned MRRs, and the ordinate represents the prediction accuracy.

Then, we demonstrate a three-layer optical backpropagation network architecture by pruning six layers of MRR banks. In general, the bits number in the simulator impacts the capabilities of the network to recognize new input data, and the performance of the optical neural network designed with equal or less than 4-bits is significantly diminished [19]. Hence, we use a 5-bit architecture and set the modulation rate to 5 GHz, and the running time of the input data is 320 ns without considering other power consumptions. Figure 7a shows the serialized input images (numbers 5, 3 and 1, respectively). Through the multiplication and accumulation of the optical architecture, the results of the MNIST dataset recognition based on the optical neural network of the pruned BP model are shown in Figure 7b. For the test of 1797 images, the overall recognition accuracy of the optical neural network is 93.49%.

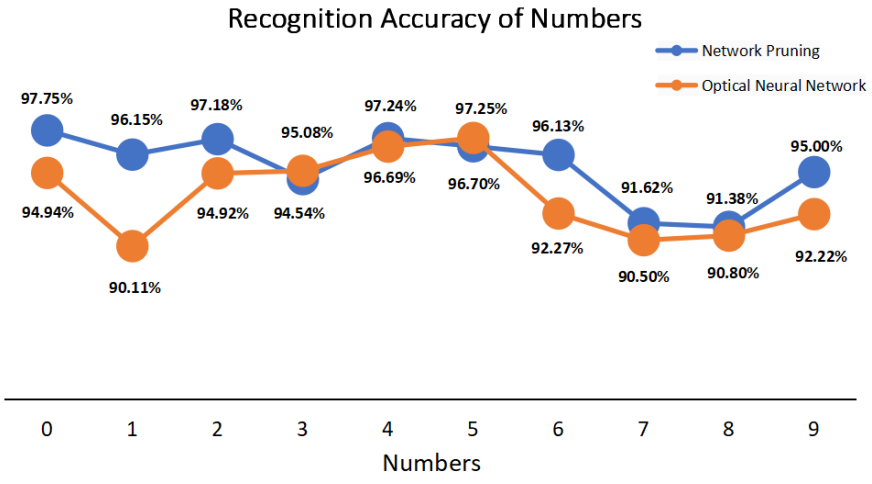
This demonstration does not realize the training process of parameters and nonlinear activation operation on the optical structure, which are also two challenges to implement for all-optical neural networks. Consequently, we will consider using nonlinear units or other optical materials to realize on-chip training and nonlinear activation. For example, using the transpose matrix operation of the MRR crossbar arrays [39] to implement the on-chip training of the optical neural networks and building an electro-optic hardware platform [40] to implement nonlinear activation functions. It is noteworthy that the optical-to-optical nonlinearity is realized by converting a small portion of the input optical signal into an analog electric signal, which is used to modulate the intensity of original optical signal with no reduction in speed of processing. This activation function is reconfigurable

via electrical bias, which allows it to be programmed or trained to synthesize a variety of nonlinear responses. Furthermore, passive optical elements [41] achieve the backpropagation gradients required in the training process by using saturable absorption for the nonlinear units. Thus, we are expected to implement a completely pruned optical BP neural network in the next work.

Future work is likely to extend to optical quantum neural networks, as many features of quantum optics can be directly mapping to neural networks [42], and technological advances driven by the trends of the photon quantum computing and optoelectronic industry provide possible venues for the large-scale and high bandwidth localization of quantum optical neural networks. Programmable silicon photonic devices can simulate the quantum walking dynamics of relevant particles, and all important parameters can be fully controlled, including the hamiltonian structure, evolution time, particle resolution and exchange symmetry [43]. Removing redundant photon devices in the universal unitary process by weight-elimination can facilitate the construction of large-scale and low-cost optical quantum neural networks.



(**a**)



(**b**)

Figure 7. (**a**) The input images are serialized. The red signal is digit 5, the black signal is digit 3 and the blue one is digit 1. (**b**) Results of the MNIST task when pruning 6 layers of MRR banks in an optical backpropagation network with an overall accuracy of 93.49%.

4. Conclusions

In this paper, we demonstrate a three-ply optical neural network based on a pruned BP model. Through pruning the model in training based on weight-elimination, an optical neural network is used to perform matrix multiplication in inference. The prediction accuracy is the best when six nodes are pruned in a hidden layer by comparing different pruned MRR banks. Furthermore, results show that the prediction accuracy of a pruned model can reach 93.49% in the MNIST dataset. In terms of energy efficiency, when pruning multiple MRR weight banks, the photonic integrated circuits become more streamlined and energy-efficient. Although the training process of parameters and nonlinear activation operation are currently incomplete on the optical structure, all of these problems are expected to be solved in the future with the development of nonlinear units and further research on optical materials, such as an MRR crossbar array [39], an electro-optic architecture for synthesizing optical-to-optical nonlinearities [40] and passive optical elements [41].

In summary, pruning different nodes has important guiding significance for optical matrix components of all-optical neural networks. Significantly, this work is scalable to much more complex networks and suitable for different optical devices, and it offers a feasible path toward truly large-scale optical neural networks.

Author Contributions: Q.Z., Z.X. and D.H. designed the study, Q.Z. and Z.X. performed the research, analysed data, and were involved in writing the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (NSFC) (Grants No. 61801522), the National Natural Science Foundation of Hunan Province, China (Grant No. 2019JJ40352).

Data Availability Statement: Publicly available datasets were analyzed in this study. The MNIST database is "Modified National Institute of Standards and Technology database". These data can be accessed and found here at any time: https://archive.ics.uci.edu/ml/machine-learning-databases/optdigits/.

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

References

1. Sallab, A.E.; Abdou, M.; Perot, E.; Yogamani, S. Deep reinforcement learning framework for autonomous driving. *Electron. Imaging* **2017**, *2017*, 70–76. [CrossRef]
2. Singh, R.; Srivastava, S. Stock prediction using deep learning. *Multimed. Tools Appl.* **2017**, *76*, 18569–18584. [CrossRef]
3. Zhang, J.; Zong, C. Deep Neural Networks in Machine Translation: An Overview. *IEEE Intell. Syst.* **2015**, *30*, 16–25. [CrossRef]
4. Lu, C.; Tang, X. Surpassing human-level face verification performance on LFW with GaussianFace. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015.
5. Rahim, A.; Spuesens, T.; Baets, R.; Bogaerts, W. Open-Access Silicon Photonics: Current Status and Emerging Initiatives. *Proc. IEEE* **2018**, *106*, 2313–2330. [CrossRef]
6. Soref, R.A.; Bennett, B.R. Electrooptical effects in silicon. *IEEE J. Quantum Electron.* **1987**, *23*, 123–129. [CrossRef]
7. Cardenas, J.; Poitras, C.B.; Robinson, J.T.; Preston, K.; Chen, L.; Lipson, M. Low loss etchless silicon photonic waveguides. *Opt. Express* **2009**, *17*, 4752–47577. [CrossRef]
8. Zhou, Z.; Chen, R.; Li, X.; Li, T. Development trends in silicon photonics for data centers. *Opt. Fiber Technol.* **2018**, *44*, 13–23. [CrossRef]
9. Tamura, P.N.; Wyant, J.C. Two-Dimensional Matrix Multiplication using Coherent Optical Techniques. *Opt. Eng.* **1979**, *18*, 182198. [CrossRef]
10. Xiang, S.; Han, Y.; Song, Z.; Guo, X.; Zhang, Y.; Ren, Z.; Wang, S.; Ma, Y.; Zou, W.; Ma, B.; et al. A review: Photonics devices, architectures, and algorithms for optical neural computing. *J. Semicond.* **2021**, *42*, 023105. [CrossRef]
11. Shen, Y.; Harris, N.C.; Skirlo, S.; Englund, D.; Soljacic, M. Deep learning with coherent nanophotonic circuits. *Nat. Photon* **2017**, *11*, 441–446. [CrossRef]
12. Tait, A.N.; Lima, T.; Zhou, E.; Wu, A.X.; Nahmias, M.A.; Shastri, B.J.; Prucnal, P.R. Neuromorphic photonic networks using silicon photonic weight banks. *Sci. Rep.* **2017**, *7*, 7430. [CrossRef]
13. Feldmann, J.; Youngblood, N.; Wright, C.D.; Bhaskaran, H.; Pernice, W. All-optical spiking neurosynaptic networks with self-learning capabilities. *Nature* **2019**, *569*, 208–214 . [CrossRef]
14. Xu, X.; Tan, M.; Corcoran, B.; Wu, J.; Boes, A.; Nguyen, T.G.; Chu, S.T.; Little, B.E.; Hicks, D.G.; Morandotti, R. 11 TOPS photonic convolutional accelerator for optical neural networks. *Nature* **2021**, *589*, 44–51. [CrossRef]

15. Xiang, S.; Ren, Z.; Song, Z.; Zhang, Y.; Hao, Y. Computing Primitive of Fully VCSEL-Based All-Optical Spiking Neural Network for Supervised Learning and Pattern Classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 2494–2505. [CrossRef] [PubMed]
16. Hughes, T.W.; Momchil, M.; Yu, S.; Fan, S. Training of photonic neural networks through in situ backpropagation. *Optica* **2018**, *5*, 864–871. [CrossRef]
17. Yang, L.; Zhang, L.; Ji, R. On-chip optical matrix-vector multiplier. *Proc. SPIE Int. Soc. Opt. Eng.* **2013**, *8855*, 88550F. [CrossRef]
18. Tait, A.N.; Wu, A.X.; Lima, T.; Zhou, E.; Prucnal, P.R. Microring Weight Banks. *IEEE J. Sel. Top. Quantum Electron.* **2016**, *22*, 1–14. [CrossRef]
19. Bangari, V.; Marquez, B.A.; Miller, H.; Tait, A.N.; Nahmias, M.A. Digital Electronics and Analog Photonics for Convolutional Neural Networks (DEAP-CNNs). *IEEE J. Sel. Top. Quantum Electron.* **2020**, *26*, 7701213. [CrossRef]
20. Lin, X.; Rivenson, Y.; Yardimci, N.T.; Veli, M.; Luo, Y.; Jarrahi, M.; Ozcan, A. All-optical machine learning using diffractive deep neural networks. *Science* **2018**, *361*, 1004–1008. [CrossRef] [PubMed]
21. Chang, J.; Sitzmann, V.; Dun, X.; Heidrich, W.; Wetzstein, G. Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification. *Nature* **2018**, *8*, 12324. [CrossRef] [PubMed]
22. Qian, C.; Lin, X.; Lin, X.; Xu, J.; Chen, H. Performing optical logic operations by a diffractive neural network. *Light. Sci. Appl.* **2020**, *9*, 59. [CrossRef]
23. Zuo, Y.; Li, B.; Zhao, Y.; Jiang, Y.; Chen, Y.C.; Chen, P.; Jo, G.B.; Liu, J.; Du, S. All Optical Neural Network with Nonlinear Activation Functions. *Optica* **2019**, *6*, 1132–1137. [CrossRef]
24. Hamerly, R.; Bernstein, L.; Sludds, A.; Soljai, M.; Englund, D. Large-Scale Optical Neural Networks Based on Photoelectric Multiplication. *Phys. Rev. X* **2019**, *9*, 021032. [CrossRef]
25. Shi, B.; Calabretta, N.; Stabile, R. Deep Neural Network Through an InP SOA-Based Photonic Integrated Cross-Connect. *IEEE J. Sel. Top. Quantum Electron.* **2020**, *26*, 1–11. [CrossRef]
26. Xu, S.; Wang, J.; Wang, R.; Chen, J.; Zou, W. High-accuracy optical convolution unit architecture for convolutional neural networks by cascaded acousto-optical modulator arrays. *Opt. Express* **2019**, *27*, 19778–19787. [CrossRef]
27. Xu, S.; Wang, J.; Zou, W. Optical Convolutional Neural Network with WDM-Based Optical Patching and Microring Weighting Banks. *IEEE Photonics Technol. Lett.* **2021**, *33*, 89–92. [CrossRef]
28. Wetzstein, G.; Ozcan, A.; Gigan, S.; Fan, S.; Englund, D.; Soljacic, M.; Denz, C.; Miller, D.A.B.; Psaltis, D. Inference in artificial intelligence with deep optics and photonics. *Nature* **2020**, *558*, 39–47. [CrossRef] [PubMed]
29. Huang, C.; Bilodeau, S.; Lima, T.; Tait, A.N.; Prucnal, P.R. Demonstration of scalable microring weight bank control for large-scale photonic integrated circuits. *APL Photonics* **2020**, *5*, 040803. [CrossRef]
30. LeCun, Y.; Denker, J.; Solla, S. Optimal brain damage. In *Advances in Neural Information Processing Systems*; ACM: New York, NY, USA, 1990; pp. 598–605. [CrossRef]
31. Hecht-Nielsen. Theory of the backpropagation neural network. *Neural Netw.* **1988**, *1*, 445. [CrossRef]
32. Zhu, M.; Gupta, S. To prune, or not to prune: Exploring the efficacy of pruning for model compression. *arXiv* **2017**, arXiv:1710.01878.
33. Weigend, A.S.; Rumelhart, D.E. Weight elimination and effective network size. In *Proceedings of a Workshop on Computational Learning Theory and Natural Learning Systems*; ACM: New York, NY, USA, 1994; Volume 1, pp. 457–476. [CrossRef]
34. Rissanen, J. Modeling by shortest data description. *Automatica* **1978**, *14*, 465–471. [CrossRef]
35. Jayatilleka, H.; Murray, K.; Caverley, M.; Jaeger, N.; Chrostowski, L.; Shekhar, S. Crosstalk in SOI Microring Resonator-Based Filters. *J. Light. Technol.* **2016**, *34*, 2886–2896. [CrossRef]
36. Zhu, S.; Liow, T.Y.; Lo, G.Q.; Kwong, D.L. Fully complementary metal-oxide-semiconductor compatible nanoplasmonic slot waveguides for silicon electronic photonic integrated circuits. *Appl. Phys. Lett.* **2011**, *98*, 83. [CrossRef]
37. Baehr-Jones, T.; Ding, R.; Ayazi, A.; Pinguet, T.; Streshinsky, M.; Harris, N.; Li, J.; He, L.; Gould, M.; Zhang, Y.; et al. A 25 Gb/s Silicon Photonics Platform. *arXiv* **2012**, arXiv:1203.0767v1.
38. Jayatilleka, H.; Murray, K.; Guillén-Torres, M.; Caverley, M.; Hu, R.; Jaeger, N.; Chrostowski, L.; Shekhar, S. Wavelength tuning and stabilization of microring-based filters using silicon in-resonator photoconductive heaters. *Opt. Express* **2015**, *23*, 25084–25097. [CrossRef]
39. Ohno, S.; Toprasertpong, K.; Takagi, S.; Takenaka, M. Si microring resonator crossbar array for on-chip inference and training of optical neural network. *arXiv* **2021**, arXiv:2106.04351v2.
40. Williamson, I.A.D.; Hughes, T.W.; Minkov, M.; Bartlett, B.; Pai, S.; Fan, S. Reprogrammable Electro-Optic Nonlinear Activation Functions for Optical Neural Networks. *IEEE J. Sel. Top. Quantum Electron.* **2020**, *26*, 7700412. [CrossRef]
41. Guo, X.; Barrett, T.D.; Wang, Z.M.; Lvovsky, A.I. Backpropagation through nonlinear units for the all-optical training of neural networks. *Photonics Res.* **2021**. *9*, 71–80. [CrossRef]
42. Steinbrecher, G.R.; Olson, J.P.; Englund, D.; Carolan, J. Quantum optical neural networks. *NPJ Quantum Inf.* **2019**, *5*, 60. [CrossRef]
43. Qiang, X.; Wang, Y.; Xue, S.; Ge, R.; Wu, J. Implementing graph-theoretic quantum algorithms on a silicon photonic quantum walk processor. *Sci. Adv.* **2021**, *7*, eabb8375. [CrossRef] [PubMed]