Hindawi
International Journal of Optics
Volume 2021, Article ID 6667495, 16 pages
https://doi.org/10.1155/2021/6667495

Research Article

Nonlinear All-Optical Diffractive Deep Neural Network with 10.6 μm Wavelength for Image Classification

Yichen Sun, Mingli Dong, Mingxin Yu, Jiabin Xia, Xu Zhang, Yuchen Bai, Lidan Lu, and Lianqing Zhu

Key Laboratory of the Ministry of Education for Optoelectronic Measurement Technology and Instrument, Beijing Information Science and Technology University, Beijing 100016, China

Correspondence should be addressed to Mingli Dong; dongml@bistu.edu.cn and Mingxin Yu; yumingxin@bistu.edu.cn

Received 18 October 2020; Accepted 13 February 2021; Published 28 February 2021

Academic Editor: Paramasivam Senthilkumaran

A photonic artificial intelligence chip is based on an optical neural network (ONN), low power consumption, low delay, and strong antiinterference ability. The all-optical diffractive deep neural network has recently demonstrated its inference capabilities on the image classification task. However, the size of the physical model does not have miniaturization and integration, and the optical nonlinearity is not incorporated into the diffraction neural network. By introducing the nonlinear characteristics of the network, complex tasks can be completed with high accuracy. In this study, a nonlinear all-optical diffraction deep neural network (N-D^{2}NN) model based on 10.6 μm wavelength is constructed by combining the ONN and complex-valued neural networks with the nonlinear activation function introduced into the structure. To be specific, the improved activation function of the rectified linear unit (ReLU), i.e., Leaky-ReLU, parametric ReLU (PReLU), and randomized ReLU (RReLU), is selected as the activation function of the N-D^{2}NN model. Through numerical simulation, it is proved that the N-D^{2}NN model based on 10.6 μm wavelength has excellent representation ability, which enables them to perform classification learning tasks of the MNIST handwritten digital dataset and Fashion-MNIST dataset well, respectively. The results show that the N-D^{2}NN model with the RReLU activation function has the highest classification accuracy of 97.86% and 89.28%, respectively. These results provide a theoretical basis for the preparation of miniaturized and integrated N-D^{2}NN model photonic artificial intelligence chips.

1. Introduction

Deep learning is a branch of machine learning that has been successfully used in various applications, such as image classification [1], natural language processing [2], and speech recognition [3]. Generally, deep neural networks have a remarkable layer, a connection with many parameters, making it highly capable of learning better feature representation [4]. Although the training phase for learning network weights can be completed on the graphic processing units (GPU), large models also require enough power and storage during inference because of millions of repeated memory references and matrix multiplication. Optical computing has high bandwidth and speed, inherently parallel processing, and low power compared with digitally implemented neural networks. A variety of methods for optical neural networks (ONN) have been proposed, including Hopfield networks with LED arrays [5], optoelectronic implementation of reservoir computing [5, 6], spiking recurrent networks with micron resonators [7, 8], and fully connected feedforward networks using Mach–Zehnder interferometers (MZIs) [9]. ONN uses optical methods to construct the neural network, which has many interconnected linear layers, and has the unique advantages of parallel processing, high-density wiring, and direct image processing. It can be realized by free-space optical interconnection (FSOI) and waveguide optical interconnection (WOI).

FSOI can be implemented ONN by a spatial light modulator (SLM), microlens arrays (MLA), and holographic element (HOE). HOE is an optical element made according to holography, which is generally formed by a photosensitive

film [10, 11]. Many researchers have explored diffractive optical element (DOE) based on the principle of diffraction. Bueno et al. introduced a network consisting of up to 2025 diffraction photonic nodes and formed a large-scale recursive photonic network. A digital micromirror device (DMD) is used to realize reinforcement learning with significant convergence results. Network consists of 2025 nonlinear network nodes, and each node is an SLM pixel. Moreover, DOE is used to implement a complex network structure [12]. Sheler Maktoobi et al. investigated diffraction coupled photonic networks with 30000 photons and described its extensibility in detail [13]. Lin et al. from UCLA realized the all-optical diffraction deep neural network (D^2NN). They moved the neural network from the chip to the real world in 2018, and the chip relies on the propagation of light and achieves almost zero consumption and zero delays in deep learning [14, 15]. The physical model consists of an input layer, 5 hidden layers, and an output layer. A terahertz band light source illuminates the input layer, and the phase or amplitude of the input surface encodes optical information. The incident light is diffracted through the input layer, and the hidden layer modulates the phase or amplitude of the light. An array of photodetectors at the output layer detects the intensity of the output light and identifies handwritten digits based on the difference in light intensity of 10 different areas. The updated phase models the diffraction grating produced by 3D printing. However, this scheme has some defects. Except for the lack of miniaturization and integration, the 3D-printed diffraction grating layer cannot be rapidly programmed in real-time. In 2019, the team proposed a wideband diffraction neural network based on the above architecture [16]. The requirements of the model for the light source are no longer limited to monochromatic coherent light, and the application scope of the framework is extended. However, the experimental environment is limited by using terahertz light sources, the large size of the diffraction grating goes against integration, and in the D^2NN model, the author stated that no activation function was added in the simulation state; so the nonlinear representation ability and generalization ability of the model need to be improved. Thus, a phase grating was used in our previous work to replace the 3D-printed diffraction grating. The carbon dioxide laser is used to emit a 10.6 μm infrared laser, and HgCdTe detection array is used to detect the light transmitted from the output layer. The size of each neuron can be reduced to 5 μm, so that a 1 mm $\times$ 1 mm phase grating can contain 200 $\times$ 200 neurons. Thus, this kind of diffraction grating will obtain a wider range of applications [17]. The advantage of this diffraction grating is that it has the size of 1 mm $\times$ 1 mm, which is conducive to miniaturization and integration of all-optical D^2NN architecture.

At present, a complex-valued neural network [18] has been successfully used for various tasks [19–27], such as processing and analysis of complex numerical data and tasks with intuitive mapping to complex numbers. Image and signal transformation in waveform or Fourier transform has been used as input data of complex numerical neural networks [28]. In the ONN, due to the complexity of the phase value of light, the phase and amplitude of light need to be widely considered. If only a real-valued neural network is used, ignoring imaginary parameters, part of the information would omit [29, 30]. Therefore, it is necessary to apply complex-valued neural networks to optical computing.

Nonlinear activation functions are widely used in various neural networks. It plays a crucial role in neural networks by learning the complex mapping between input and output. If there is no activation function in the neural network and no matter how many neural networks there are, the output is a linear combination of inputs. This means that the system lacks a hidden layer, resulting in a low nonlinear representation ability of the model. At present, nonlinear activation functions mainly include sigmoid, tanh, and ReLU. Thereinto, ReLU is the most common ones for three reasons: (1) solving the so-called explosion and gradient disappearance, (2) accelerating convergence [31], and (3) making the output of some neurons 0, which leads to the sparse network. ReLU activation function includes Leaky-ReLU, PReLU, and RReLU. These functions improve the speed and accuracy of classifying different datasets. ReLU activation function allows the network itself to introduce sparsity. This method is equivalent to the pretraining of unsupervised learning and greatly shortens the learning cycle.

In this study, an all-optical diffraction deep neural network (N-D^2NN) model with nonlinear activation functions based on a 10.6 μm wavelength is proposed. Comparing with the work investigated by UCLA [14, 15], the characteristic size of the neural network is reduced by 80 times, and the classification accuracy of the model is verified by simulation. Our model provides a theoretical basis for the future research of the N-D^2NN model framework in 10.6 μm wavelength and lays a foundation for the further realization of large-scale integrated and miniaturized photonic computation chips.

In summary, the main contributions of this study are as follows: (1) an N-D^2NN framework with nonlinear activation functions based on 10.6 μm wavelength is proposed by combining ONN and complex-valued neural networks. (2) The representation ability of N-D^2NN with ReLU improvement activation functions is evaluated in the experimental simulation state, and the detailed evaluation process is given.

The rest of this study is organized as follows. The method used in our research is described in Section 2. Section 3 presents the experimental results. The discussion is reported in Section 4. Finally, conclusions are given.

2. Materials and Methods

This part introduces the basic theory and improved diffraction deep neural network method based on a 10.6 μm laser wavelength. First, the optical calculation theory of N-D^2NN based on 10.6 μm wavelength is introduced. Then, the network model structure is explained in detail. Finally, to improve the nonlinear representation ability of N-D^2NN, an improved method of N-D^2NN is given by adding the nonlinear activation function into the N-D^2NN model.

2.1. Optical Computation. Figure 1 shows the structure of N-D^{2}NN. Light passing through each grating is modulated by grating grids of different thickness, and it is then received by all grating pixels on the secondary grating. This network connection mode is similar to the fully connected neural network. The first layer of grating receives input images and corresponds to the input layer in the neural network structure. The middle layers of gratings correspond to the hidden layers in the neural network structure, and the detection plane corresponds to the output layer in the neural network structure. The phase modulation effect of the input light is different from the height of different gratings, which corresponds to different weights in the neural network structure.

According to the Rayleigh–Sommerfeld diffraction equation, the neurons in each layer of N-D^{2}NN can be calculated by the secondary wave source equation, and the formula is as follows [32, 33]:

$$w_i^l(x, y, z) = \frac{z - z_i}{r_2}\left(\frac{1}{2\pi r} + \frac{1}{j\lambda}\right)\exp\left(\frac{j2\pi r}{\lambda}\right), \tag{1}$$

where l represents the l^{th} layer of the network, i represents the i^{th} neuron of layer l, r represents the Euclidean distance between l layer node i and $l+1$ layer node, and $j = \sqrt{-1}$. The input plane is the 0^{th} layer, and then, for l^{th} layer ($l \geq 1$), the output field can be expressed as

$$n_i^l(x, y, z) = w_i^l(x, y, z) \cdot g, \tag{2}$$

where $n_i^l(x, y, z)$ represents the output of the i^{th} neuron at the l^{th} layer (x, y, z), g represents the nonlinear activation function in the neural network whose function is to transmit the modulated second-wave neurons to the next layer through the nonlinear unit, and $g = \phi[t_i^l(x_i, y_i, z_i) \cdot \sum_k n_k^l - 1(x_i, y_i, z_i)] = \phi[w_i^l(x, y, z) \cdot |A| \cdot e^{j\phi_i^l}]$. t_i^l denotes the complex modulation, i.e., $t_i^l(x_i, y_i, z_i) = |A|\exp(j\phi_i^l(x_i, y_i, z_i))$, $|A| = a_i^l(x_i, y_i, z_i)$ is the relative amplitude of the secondary wave, and $\phi_i^l(x_i, y_i, z_i)$ represents the phase delay increased by the input wave $\sum_k n_k^{l-1}(x_i, y_i, z_i)$ and the complex-valued neuron modulation function t_i^l on each neuron. For N-D^{2}NN structure with the only phase, the amplitude $a_i^l(x_i, y_i, z_i)$ is considered a constant, and the ideal state is 1 when the optical loss is ignored.

2.2. The Architecture of N-D^{2}NN. To simplify the representation of the forward model, equation (1) can be rewritten as

$$\begin{cases} n_{i,p}^l = w_{i,p}^l \cdot g, \\ m_i^l = \sum_k n_k^{l-1}, \\ t_i^l = a_i^l \exp\left(j\phi_i^l\right), \\ g = \phi\left(m_i^l \cdot t_i^l\right), \end{cases} \tag{3}$$

where i refers to a neuron of the l^{th} layer, and p refers to a neuron of the next layer, connected to neuron i by optical diffraction. The input pattern h_k^0 is located at layer 0. It generally has a complex-valued quantity, which can carry information in its phase and amplitude channels. The diffraction wave function generated by the interaction between illumination plane wave and input light can be expressed as

$$n_{k,p}^0 = w_{k,p}^0 \cdot h_k^0. \tag{4}$$

When the input light is diffracted through a multilayer grating, a result image will be output on the detection plane. The detector detects the detection area in the generated image and obtains the network classification result. Therefore, it is necessary to process the data labels in the parameter training stage, and the corresponding labels are designed in the resulting images of different labels. As shown in Figure 2, by judging the region with the highest light intensity in the detection region of the generated image, the label represented by the generated image can be obtained. To match input data of different lengths, the resulting image corresponding to the label is also scaled.

After the input light is diffracted by multilayer grating, a result image will be output in the detection plane. The detector probes the detection area in the resulting image to obtain the network classification results. Therefore, it is necessary to process the data labels in the parameter training stage and design different labels to correspond to the marks in the resulting image, as shown in Figure 2. The label represented by the resulting image can be obtained by judging the region with the highest light intensity in the detection region of the resulting image. The resulting image corresponding to the label needs to be scaled to match input data of different lengths.

For N-D^{2}NN containing N hidden layers, the light intensity of its output layer can be expressed as

$$I_i^{N+1} = \left|m_i^{N+1}\right|^2. \tag{5}$$

The intensity measured by the detector on the output plane is normalized so that they are located in the interval (0, 9) of each sample. I_l is used to represent the total amount of optical signals incident on the detector in the output layer l, and the normalized intensity I_l' is

$$I_l' = \frac{I_l}{\max\{I_l\}} \times 10. \tag{6}$$

2.3. The Proposed Method. Based on a previous research, Lin et al. did not consider adding nonlinearity to the D^{2}NN framework. Therefore, in the classification task, D^{2}NN is weak in nonlinear representation. In this study, an N-D^{2}NN model architecture is proposed, as shown in Figure 3. Assume that a neuron is physically equivalent to a grid of ONN, and the modulated secondary wave neurons are transmitted to the next layer through the nonlinear unit, as shown in Figure 3.

2.3.1. Complex-Valued Neural Network. According to equation (3), the phase factor in the complex form of the wave function contains the spatial phase factor $\exp(j\phi_i^l)$, so

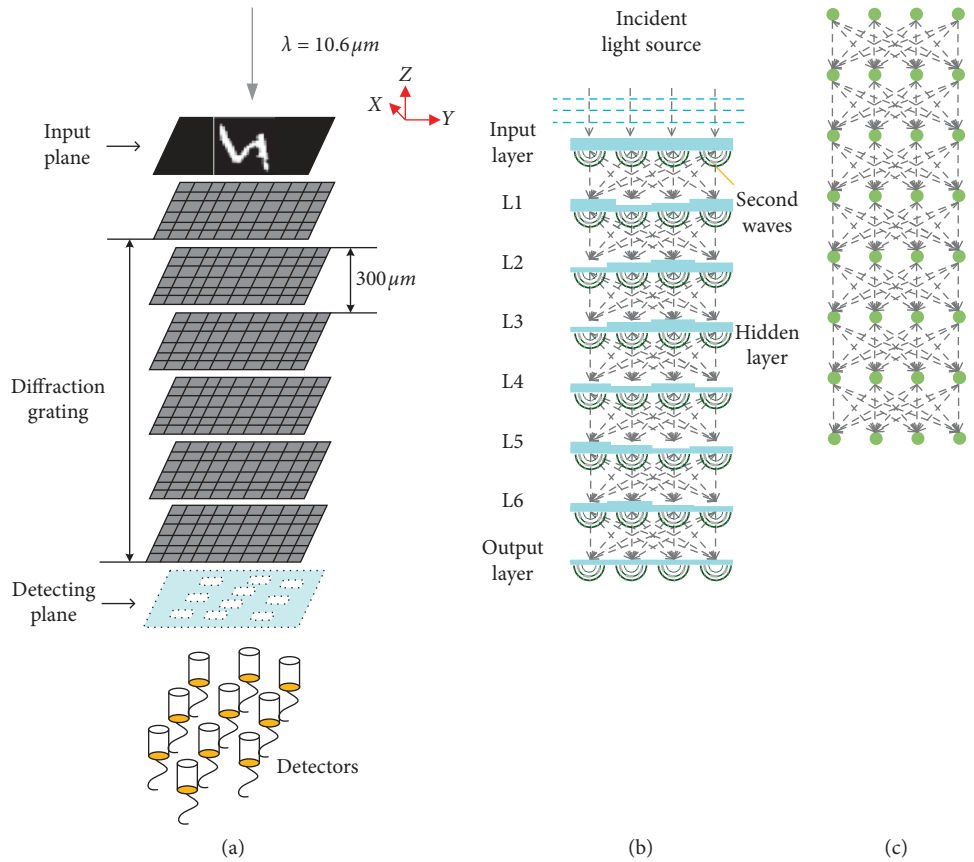

Figure 1: Schematic diagram of the N-D^{2}NN structure. (a) System physical model. (b) Optical path model. (c) Neural network model.

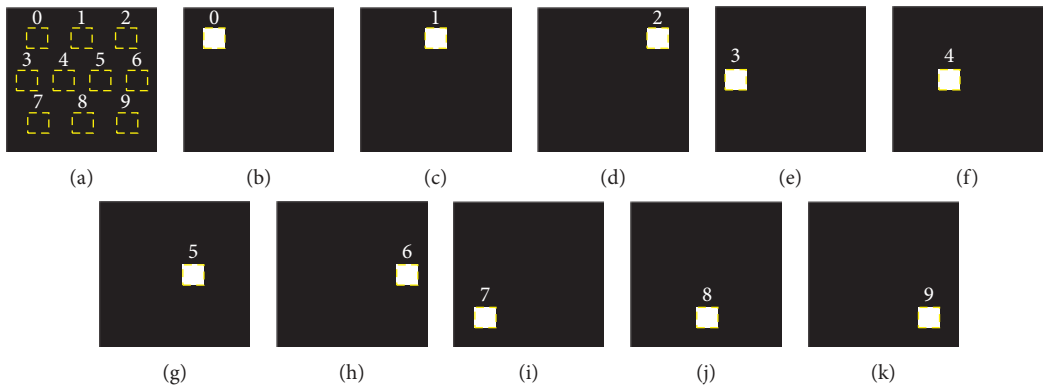

Figure 2: Image label design. (a) Detection area. (b) Label 0. (c) Label 1. (d) Label 2. (e) Label 3. (f) Label 4. (g) Label 5. (h) Label 6. (i) Label 7. (j) Label 8. (k) Label 9.

the product of the amplitude and the spatial phase factor is $t_i^l = x + jy = a_i^l \exp(j\phi_i^l)$. t_i^l can be represented by two real numbers: the real part $\mathrm{Re}(t_i^l) = x$, and the imaginary part $\mathrm{Im}(t_i^l) = y$. Any complex-valued function of multiple complex variables can be represented by two functions: $f(t_i^l) = f(x, y) = f(a_i^l, \phi_i^l)$.

Although directly used and represented in neural networks, complex numbers define the interaction between two

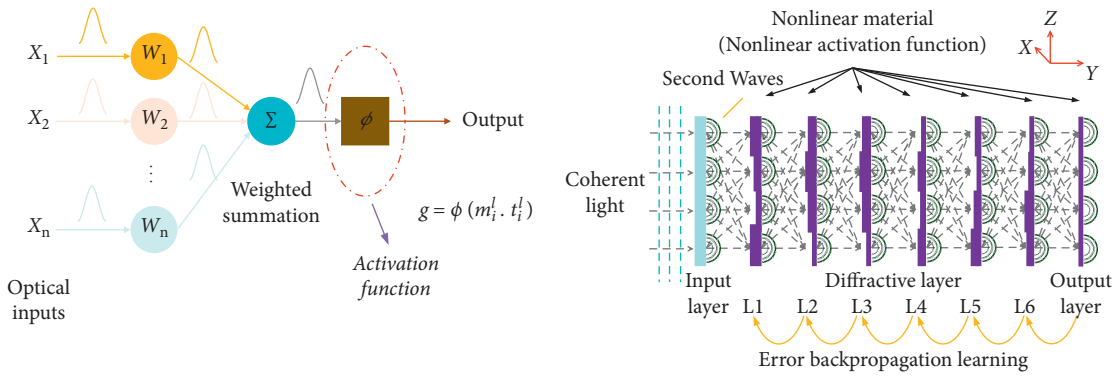

FIGURE 3: (a) Schematic diagram of neurons in the diffraction mode with nonlinear activation. (b) N-D^{2}NN blueprints with optical nonlinear materials.

parts. Using Euler's constant $e^{j\phi_i^l} = \cos(\phi_i^l) + j\sin(\phi_i^l)$ as the equivalent representation in the form of polarity,

$$t_{1i}^l t_{2i}^l = \left(a_{1i}^l e^{-j\phi_{1_i^l}}\right)\left(a_{2i}^l e^{-j\phi_{2_i^l}}\right), \tag{7}$$

$$\begin{aligned} t_{1i}^l + t_{2i}^l = &\left(a_{1i}^l \cos\left(\phi_{1_i^l}\right) + a_{2i}^l \cos\left(\phi_{2_i^l}\right)\right. \\ &\left. + j\left(a_{2i}^l \sin\left(\phi_{2_i^l}\right) + a_{1i}^l \sin\left(\phi_{1_i^l}\right)\right)\right). \end{aligned} \tag{8}$$

Because more operations are required, complex parameters increase the complexity of the neural network. Therefore, equations (7) and (8) can be used according to the selected implementation mode and representation, which can significantly reduce the computational complexity. The product of input t_i^l and complex numerical weight matrix w_i^l is calculated as follows:

$$t_i^l w_i^l = \begin{bmatrix} \mathrm{Re}\left(t_i^l\right) & -\mathrm{Im}\left(t_i^l\right) \\ \mathrm{Im}\left(t_i^l\right) & \mathrm{Re}\left(t_i^l\right) \end{bmatrix} \begin{bmatrix} \mathrm{Re}\left(w_i^l\right) \\ \mathrm{Im}\left(w_i^l\right) \end{bmatrix} = \begin{bmatrix} \mathrm{Re}\left(t_i^l\right)\mathrm{Re}\left(w_i^l\right) - \mathrm{Im}\left(t_i^l\right)\mathrm{Im}\left(w_i^l\right) \\ \mathrm{Im}\left(t_i^l\right)\mathrm{Re}\left(\left(w_i^l\right) + \mathrm{Re}\left(t_i^l\right)\mathrm{Im}\left(w_i^l\right) \end{bmatrix}. \tag{9}$$

So this exchange means that the model design needs to be rethought to simplify the structure. A deep learning framework that performs poorly under real-valued parameters may be suitable for complex-valued parameters. According to the experimental results in [34], real-valued data do not require this structure. The imaginary part of $\mathrm{Im}(t_i^l)$ is zero, so equation (9) can be simplified as

$$\begin{aligned} \mathrm{Re}\left(t_i^l w_i^l\right) &= \mathrm{Re}\left(t_i^l\right)\mathrm{Re}\left(w_i^l\right), \\ \mathrm{Im}\left(t_i^l w_i^l\right) &= \mathrm{Re}\left(t_i^l\right)\mathrm{Im}\left(w_i^l\right). \end{aligned} \tag{10}$$

For training, this means that the real parts $\mathrm{Re}(t_i^l)$ and $\mathrm{Re}(w_i^l)$ dominate the overall classification of the real-valued data points.

2.3.2. Activation Function. The activation function can enhance the representation ability of nonlinearity and perform a complex task of deep learning. However, in some nonlinear activation functions, such as sigmoid and tanh, they have two disadvantages: (1) when performing backpropagation to calculate the error gradient and calculating the activation function (exponential function), the derivation involves division, so the computation is relatively large, and (2) when the sigmoid is close to the saturation region, the transformation is too slow, and the derivative tends to zero. This situation will cause information loss. In all of these nonlinear activation functions, the most notable one is the rectified linear unit (ReLU) [35]. It is generally believed that the excellent performance of ReLU comes from sparsity [36, 37]. It reduces the interdependence of parameters and alleviates the occurrence of overfitting problems. There are also some improvements to ReLU, such as leaky rectified linear (Leaky-ReLU), parametric rectified linear (PReLU), and randomized rectified linear (RReLU), namely, ReLU family functions. These ReLU family functions improve the speed and accuracy of neural network training. In this section, the three kinds of rectified units are introduced: Leaky-ReLU, PReLU, and RReLU. They are illustrated in Figure 4.

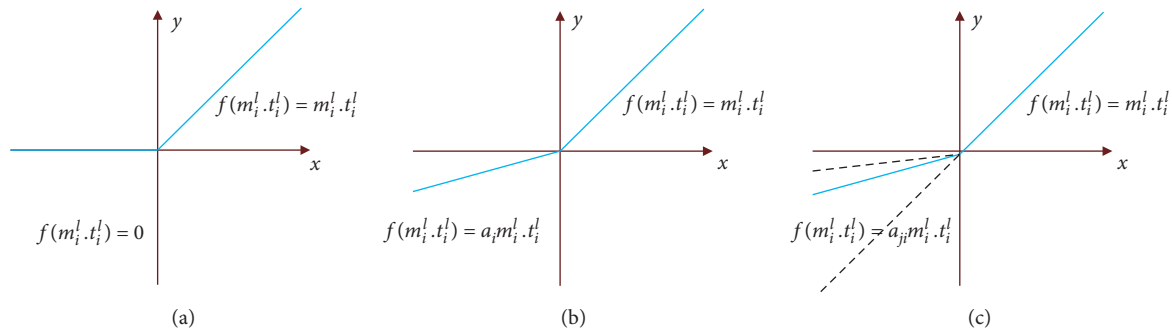

Figure 4: Mathematical models of (a) ReLU, (b) Leaky-ReLU/PReLU, and (c) RReLU functions.

Figure 4(a) shows the mathematical model of ReLU, which is first used in restricted Boltzmann machines. It is a piecewise linear function that cuts the negative part to zero and keeps the positive part. After passing with ReLU, activation is sparse. Formally, rectified linear activation is defined as

$$f\left(m_i^l \cdot t_i^l\right) = \begin{cases} m_i^l \cdot t_i^l, & \text{if } m_i^l \cdot t_i^l \geq 0, \\ 0, & \text{if } m_i^l \cdot t_i^l < 0, \end{cases} \tag{11}$$

where input signal $m_i^l \cdot t_i^l < 0$ and output is 0; when the input signal $m_i^l \cdot t_i^l \geq 0$, the output is equal to the input signal.

Figure 4(b) shows the mathematical model of Leaky-ReLU and PReLU. ReLU sets all negative values to zero. In contrast, leaky rectified linear unit (Leaky-ReLU) assigns a nonzero slope to all negative values. Leaky-ReLU activation function is first proposed in the acoustic model [38]. It is mathematically defined as

$$f\left(m_i^l \cdot t_i^l\right) = \begin{cases} m_i^l \cdot t_i^l, & \text{if } m_i^l \cdot t_i^l \geq 0, \\ a_i m_i^l \cdot t_i^l, & \text{if } m_i^l \cdot t_i^l < 0, \end{cases} \tag{12}$$

where a_i is a fixed parameter in range (0, 1). In this study, a_i in the Leaky-ReLU function is selected as 0.2.

PReLU is proposed by He et al. [39]. The authors reported that its performance is much better than ReLU in large-scale image classification tasks. In the PReLU function, the slopes of the negative part are learned from the data rather than defined in advance. PReLU function learns a_i through back propagation during training in equation (12).

Figure 4(c) shows the mathematical model of RReLU, which is the randomized version of Leaky-ReLU. It is first proposed and used in the Kaggle NDSB competition. The highlight of RReLU is that in the training process, a_{ji} is a random number sampled from a uniform distribution $U(l, u)$. The mathematical terms are defined as

$$f\left(m_i^l \cdot t_i^l\right) = \begin{cases} m_i^l \cdot t_i^l, & \text{if } m_i^l \cdot t_i^l \geq 0, \\ a_{ji} m_i^l \cdot t_i^l, & \text{if } m_i^l \cdot t_i^l < 0, \end{cases} \tag{13}$$

where a_{ji} is an arbitrary constant in the interval $U(l, u)$, $l < u$, and $l, u \in [0, 1)$. Suggested by the NDSB competition winner, a_{ji} is sampled from U (3, 8). In this study, the same configuration is used.

2.3.3. Model Training. The forward propagation model compares the result of the physical output plane with the training target of the diffraction network, and the error propagation generated is updated iteratively to each layer of the diffraction network. Based on the reports [15], the cross-entropy function is adopted as the loss function for N-D^{2}NN, which significantly improves the classification accuracy of the MNIST dataset [40] and Fashion-MNIST dataset [41], respectively. The output results of N-D^{2}NN are compared with the input values. The error backpropagation is used to iterate the grating parameters, and the loss function is defined according to the output of N-D^{2}NN based on the target characteristics. The cross-entropy function is used as the loss function in the neural network. According to the following formula, define the cross-entropy function as

$$H(p, q) = -\sum_{l}^{K} p_i^l(x) \log q_i^l(x), \tag{14}$$

where $p_i^l(x) = e^{I'} / \sum_l^K e^{I'}$ represents the output value of the Softmax layer in the neural network, and Softmax regression can be thought of as a learning algorithm to optimize classification results. $q_i^l(x)$ represents the actual image output value, and $e^{I'}$ represents the normalized intensity of the output plane. To train the N-D^{2}NN model into a digital classifier, the MNIST handwritten digital dataset and Fashion-MNIST dataset are used as the input layers.

In Figures 5(a) and 5(b) are, respectively, the grayscale and RGB images of the diffraction grating height distribution of each layer after the training of the MNIST dataset in the simulation state, and (c) and (d) represent the output grayscale images and RGB images of each layer of the diffraction grating. To judge accuracy of the resulting image, the influence of the detection area on the background information should be removed first. Then, to obtain the prediction label, the detection area template is used to extract the resulting image. After the incident light passes through the input grating and grating layer L1-L6, the region with the largest light intensity in the final grating result image is consistent with the location of the detection area label 7 in Figures 5(c) and 5(d). In Figures 5(e) and 5(f) are, respectively, the grayscale and RGB images of the diffraction grating height distribution of each layer after the training of the Fashion-MNIST dataset in the simulation state, and (g) and (h) represent the output grayscale images and RGB

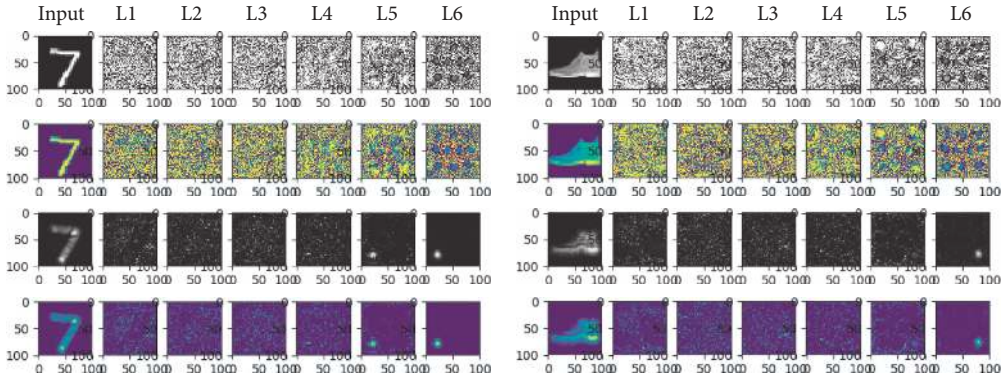

Figure 5: The height distribution images of each layer of diffraction grating and the output images of each layer of diffraction grating are obtained. (a) Label 7 in the MNIST dataset. (b) Label 9 in the Fashion-MNIST dataset.

images of each layer of the diffraction grating. After the incident light passed through the input grating and grating layer L1-L6, the area with the highest light intensity in the final grating result image was consistent with the position of the detected area label 9 (ankle boot) in Figures 5(g) and 5(h).

N-D^{2}NN was performed using the Python (3.6.4) and TensorFlow (v1.10.0, Google Inc.) framework. This model was trained on a desktop computer with a GeForce GTX TITAN V graphical processing unit (GPU) and Intel (R) Core (TM) i7-8700K CPU at 3.70 GHz and 64 GB of RAM, running Windows 10 operating system (Microsoft). The training time and the inference time of the N-D^{2}NN model using three RELU activation functions on the MNIST dataset and Fashion-MNIST dataset are shown in Tables 1 and 2, respectively. From Tables 1 and 2, it can be seen that the N-D^{2}NN model with the RReLU function takes the least training time and inference time compared with other activation functions on the MNIST and Fashion-MNIST datasets. In the training phase, the model with Leaky-ReLU and PReLU achieves the same training time on the datasets. However, the inference time of the model with Leaky-ReLU is faster than the one with PReLU. In the Kaggle NDSB competition, it is reported that a_{ji} in the RReLU function is favorable due to its randomness in training, and overfitting can be reduced. Therefore, no matter in reasoning time, training time, or recognition accuracy, RReLU function has advantages. The a_i in the Leaky-ReLU function is fixed, and the a_i in the PReLU function changes based on the data; thus, the inference time of the PReLU function is slightly longer than that of the Leaky-ReLU function.

3. Experimental Results

To test the performance of the N-D^{2}NN structure, the MNIST dataset and Fashion-MNIST dataset are introduced in Section 3.1. Section 3.2 shows the evaluation method. Performance evaluation is reported in Section 3.3. Section 3.4 discusses the comparison with the representation ability results of a neural network framework without nonlinear activation functions.

Table 1: The training time of the N-D^{2}NN model using three ReLU activation functions for the MNIST dataset and Fashion-MNIST dataset.

	Training time (h)	
	MNIST	Fashion-MNIST
Leaky-ReLU	28.1	28.2
PReLU	28.1	28.2
RReLU	27.9	28.0

Table 2: The inference time of the N-D^{2}NN model using three ReLU activation functions for the MNIST dataset and fashion-MNIST dataset.

	Inference time (s)	
	MNIST	Fashion-MNIST
Leaky-ReLU	0.12	0.14
PReLU	0.13	0.16
RReLU	0.11	0.13

3.1. MNIST Dataset and Fashion-MNIST Dataset. In this study, the MNIST handwritten digital dataset and Fashion-MNIST dataset are used as the training digital classifier at the input layer based on the 10.6 μm N-D^{2}NN model. The MNIST dataset is a handwritten digital dataset composed of numbers 0–9. The dataset comprises four parts: training set image, training set label, test set image, and test set label. The MNIST dataset comes from the National Institute of Standards and Technology (NIST). The training and testing sets are a mixture of handwritten numbers from two databases, one from high school students and the other from the Census Bureau. The MNIST handwritten dataset contains a training set of 60,000 samples and a test set of 10,000 samples. Each image in the MNIST dataset contains 28×28 pixels, and these numbers are normalized and fixed in the center.

The Fashion-MNIST dataset is a ten-category clothing dataset that replaces the MNIST handwritten number dataset. It has the same number of training sets, test sets, and

image resolutions as the MNIST dataset. However, different from the MNIST dataset, the Fashion-MNIST dataset is no longer an abstract number symbol, but a more specific clothing type. Each training sample and test sample in the MNIST dataset and Fashion-MNIST dataset are labelled according to the category in Table 3.

3.2. Evaluation Method. The confusion matrix with ten classes is listed in Table 4. First, each category H_i ($i = 0$–9) needs to compute ten in one confusion matrix [42]. Then, for a single class, the evaluation method is defined by TP_i, FN_i, TN_i, and FP_i. The following formula can express accuracy of the proposed classifier:

$$\mathrm{Accuracy} = \frac{\mathrm{TP}_i + \mathrm{TN}_i}{\mathrm{TP}_i + \mathrm{TN}_i + \mathrm{FP}_i + \mathrm{FN}_i}, \tag{15}$$

where $\mathrm{TP}_i = \chi_{ii}$ represents the totality of the predicted sample is true, and the true sample is true for H_i; $\mathrm{TN}_i = \sum_{\substack{j=0 \\ j\neq i}}^{9} \sum_{\substack{k=0 \\ k\neq i}}^{9} \chi_{jk}$ represents the totality of the predicted sample is false, and the true sample is false for H_i; $\mathrm{FP}_i = \sum_{\substack{j=0 \\ j\neq i}}^{9} \chi_{ji}$ represents the totality of the predicted sample is true and the true sample is false for H_i; and $\mathrm{FN}_i = \sum_{\substack{j=0 \\ j\neq i}}^{9} \chi_{ij}$ represents the totality of the predicted sample is false, and the true sample is true for H_i, and the totality of test samples is represented by N.

3.3. Performance Evaluation. In this study, the hyperparameters in the N-D^{2}NN model based on 10.6 μm wavelength are selected, as shown in Tables 5 and 6.

The grid search method is used to select the hyperparameters of the neural network, so the number of grating layers belongs to the hyperparameters of the neural network. In the simulation state, each batch of data in the network model is selected to be 100. To reduce the simulation time, the number of cycles is 10, the pixel scale is 28×28, the loss function is the cross-entropy function, and the optimizer is the Adam optimizer, and the learning rate is chosen as 0.01.

The number of grating layers in N-D^{2}NN based on 10.6 μm wavelength will influence the final classification result, which is also the unique advantage of this neural network compared with other linear networks. Figure 6 shows the recognition accuracy of different grating layers in N-D^{2}NN models with various activation functions. When the number of grating layers is ≤5, the classification accuracy of the neural network model increases with the number of grating layers. When the number of grating layers is >5, the classification accuracy reaches saturation. In general, the deeper the neural network is, the stronger its feature representation ability will be. Furthermore, the neural network could have a better performance on the image classification task. However, the selection of the layer number of the neural network also largely depends on the dimension of the input data features. If the feature dimension of the input data is low and the layer number of the neural network deeper, it is easy to cause the loss and saturation of the feature information during the training process. Therefore, its classification accuracy tends to be saturated or even decreased. Therefore, in the simulation experiment environment, the number of grating layers is selected as 6.

After determining the number of grating layers in the neural network model, the pixel scale and the spacing of diffraction gratings in the hyperparameters of the model are optimized, among which the number of grating layers is 6. In the N-D^{2}NN model, pixel sizes and classification accuracy corresponding to the three activation functions, Leaky-ReLU, PReLU, and RReLU, are shown in Tables 7–10, respectively.

Table 3: Label number and category of the MNIST dataset and fashion-MNIST dataset.

Label number	MNIST dataset category	Fashion-MNIST dataset category
0	0	T-shirt
1	1	Trousers
2	2	Pullover
3	3	Dress
4	4	Coats
5	5	Sandal
6	6	Shirt
7	7	Sneaker
8	8	Bag
9	9	Ankle boot

Table 4: Confusion matrix of ten-class classification.

		Predicted					
		0	1	2	...	8	9
True	0	χ_{00}	χ_{01}	χ_{02}	⋯	χ_{08}	χ_{09}
	1	χ_{10}	χ_{11}	χ_{12}	⋯	χ_{18}	χ_{19}
	2	χ_{20}	χ_{21}	χ_{22}	⋯	χ_{28}	χ_{29}
	...	⋯	⋯	⋯	⋯	⋯	⋯
	8	χ_{80}	χ_{81}	χ_{82}	⋯	χ_{88}	χ_{89}
	9	χ_{90}	χ_{91}	χ_{92}	⋯	χ_{89}	χ_{99}

Table 5: Physical parameters of neural network grating.

Grating parameter	Numerical
Wavelength	10.6 μm
Cell size	5 μm
Grating spacing	30 λ

Table 6: Neural network training parameters.

Training parameter	Numerical
Grating layer	6
Number of neurons per layer	100×100
Batch size	100
Epoch	50
Learning rate	0.05

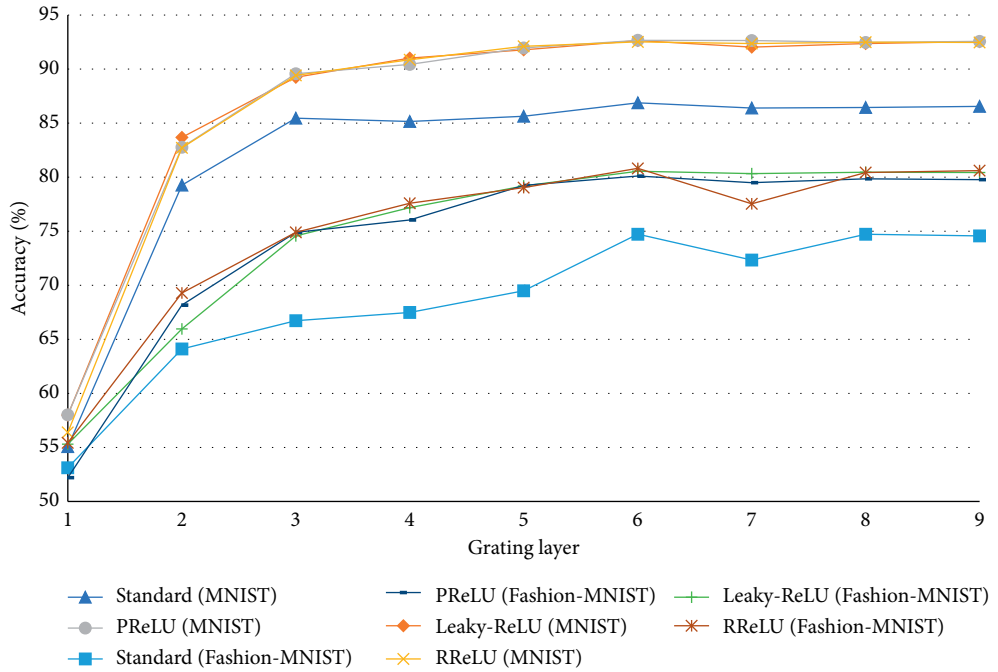

Figure 6: Classification accuracy corresponding to the number of raster layers with standard, Leaky-ReLU, PReLU, and RReLU activation function models.

Table 7: The classification accuracy in the N-D^{2}NN model adding the Leaky-ReLU activation function corresponds to pixel size and diffraction grating spacing.

Accuracy (%)		Pixel size							
		30 × 30	40 × 40	50 × 50	60 × 60	70 × 70	80 × 80	90 × 90	100 × 100
Spacing (λ)	30	93.16	94.98	95.20	95.83	96.30	96.01	96.51	96.58
	40	90.40	94.13	95.39	95.86	95.95	95.96	96.35	96.55
	50	80.79	93.74	95.04	95.64	95.79	95.98	96.35	96.55
	60	77.74	92.52	94.00	95.51	95.63	95.83	96.21	96.44
	70	68.84	89.87	93.73	95.16	95.42	96.03	96.10	96.25

Table 8: The classification accuracy in the N-D^{2}NN model adding the PReLU activation function corresponds to pixel size and diffraction grating spacing.

Accuracy (%)		Pixel size							
		30 × 30	40 × 40	50 × 50	60 × 60	70 × 70	80 × 80	90 × 90	100 × 100
Spacing (λ)	30	92.54	94.97	95.41	95.71	95.92	96.25	96.55	96.67
	40	90.02	94.51	95.02	95.67	95.88	96.23	96.41	96.44
	50	86.49	93.46	94.67	95.64	95.69	95.93	96.21	96.46
	60	77.18	92.72	94.43	95.42	95.38	96.12	96.16	96.48
	70	69.05	90.49	94.11	95.06	95.69	95.77	96.00	96.19

Table 9: The classification accuracy in the N-D^{2}NN model adding the RReLU activation function corresponds to pixel size and diffraction grating spacing.

Accuracy (%)		Pixel size							
		30 × 30	40 × 40	50 × 50	60 × 60	70 × 70	80 × 80	90 × 90	100 × 100
Spacing (λ)	30	93.10	94.93	95.08	95.81	96.06	96.15	96.35	96.78
	40	90.45	94.16	95.19	95.81	95.71	96.05	96.43	96.47
	50	85.01	93.78	95.12	95.39	95.90	95.99	96.28	96.23
	60	84.44	92.85	94.43	95.24	95.59	96.06	96.14	96.35
	70	68.88	91.27	93.71	95.03	95.49	95.67	95.85	96.08

Table 10: The classification accuracy in the standard N-D^{2}NN model corresponds to pixel size and diffraction grating spacing.

Accuracy (%)		Pixel size							
		30×30	40×40	50×50	60×60	70×70	80×80	90×90	100×100
Spacing (λ)	30	84.27	86.42	87.36	86.94	86.64	86.56	86.52	86.50
	40	82.18	86.23	87.07	87.44	86.67	86.61	86.94	86.77
	50	75.45	86.14	85.94	87.59	87.12	86.72	86.83	87.03
	60	64.99	83.27	87.10	87.14	87.41	86.94	86.80	86.97
	70	61.90	83.04	86.83	87.65	86.96	86.88	86.75	86.59

As can be seen from Tables 5–8, when the spacing of diffraction gratings in the neural network model is fixed, accuracy generally increases with pixel size. When the pixel size of the diffraction grating in the neural network model is fixed, its precision generally decreases with the increase of the spacing of the diffraction grating. When the model selects RReLU activation function, the pixel size is 100×100, and the spacing of diffraction gratings is 30λ; the neural network has the highest recognition accuracy.

Finally, the learning rate of the Adam optimizer in the model is optimized. Figure 7 shows the classification accuracy of the N-D^{2}NN model with RReLU added to the MNIST dataset. Among them, the selection learning rate is 0.01, 0.025, 0.05, and 0.075. It can be seen from Figure 7 that the classification accuracy of the model is the highest when the learning rate is 0.05.

The selected hyperparameters of the Fashion-MNIST dataset evaluated by the N-D^{2}NN model are optimized by the above method, and the selected hyperparameters are consistent with the models in the MNIST dataset. The activation function is not added into the standard N-D^{2}NN model based on 10.6 μm wavelength, and the classification accuracy of the MNIST (Fashion-MNIST) dataset obtained under the simulation state is 86.78% (81.10%).

As shown in Figure 8(a), the classification accuracy of the standard N-D^{2}NN model for each label in the MNIST dataset is not the same, and the classification accuracy of the model for label 1 is as high as 98%. However, the classification accuracy of the model to label 8 is only 73%. In Figure 8(b), the classification accuracy of the standard N-D^{2}NN model for each number in the Fashion-MNIST dataset is not the same, and the classification accuracy of the model for label 8 is as high as 95%. However, the classification accuracy of the model to label 6 is only 35%. It can be seen that the nonlinear fitting ability and generalization ability of the standard N-D^{2}NN model without the activation function is weak. According to the accuracy curve, when the epoch is 50, the accuracy of model recognition tended to be saturated.

3.4. Comparison with the N-D^{2}NN Framework. Comparison with the test results of the N-D^{2}NN structure with ReLU family nonlinear activation functions is presented in Section 3.3. Experimental simulation results show that N-D^{2}NN frameworks with different nonlinear activation functions have significantly improved representation ability. The necessity of nonlinear activation function in the N-D^{2}NN framework is proved. Leaky-ReLU, PReLU, and RReLU functions are selected as the activation functions in the N-D^{2}NN model. The classification accuracy results of the MNIST dataset and Fashion-MNIST dataset obtained under simulation are shown in Table 11.

Among them, the neural network with the RReLU function for the MNIST dataset has a classification accuracy of 97.86%. Comparing with the results shown in the [14, 15], the classification accuracy of the N-D^{2}NN model based on 10.6 μm is improved by 0.05%. The neural network with PReLU and RReLU function for the Fashion-MNIST dataset has a classification accuracy of 89.28%. This theory proves the correctness of introducing ReLU family activation functions into the model. Figure 9 shows the accuracy and confusion matrix images of N-D^{2}NN with different activation functions.

According to the accuracy image, when epoch is 50 in the model, the recognition accuracy region of the model is saturated. Confusion matrix reveals that the classification accuracy of each label in the MNIST dataset of the neural network with three activation functions is above 94%. Among them, the recognition accuracy of the model with three activation functions to the label 0 and the label 1 is as high as 99%. However, the classification ability of the model to the label 9 is slightly worse, with accuracy rates of 94%, 97%, and 94%. This may be due to the high similarity between label 9, label 4, and label 8, so the model misclassified label 9 into other labels. Figure 10 shows the recognition accuracy rate of various neural network models to various labels in the MNIST dataset. It can be seen that in the MNIST dataset, the recognition accuracy for each label of the model with three ReLU family activation functions is higher than that of the standard model without activation function.

According to the accuracy image, when epoch is 50 in the model, the recognition accuracy region of the model is also saturated. Confusion matrix reveals that the classification accuracy of each label in the Fashion-MNIST dataset of the neural network with three activation functions is above 80%, except for label 4 and label 6. Among them, the recognition accuracy of the model with three activation functions to the label 8 is as high as 98%, 96%, and 97%, respectively. However, the classification ability of the model to the label 6 is slightly worse, with accuracy rates of 58%, 66%, and 62%, respectively. The low recognition accuracy of the model for label 6 (shirt) may be because it is mistakenly divided into label 0 (T-shirt), label 2 (pullover), and label 4 (coat). Figure 11 shows the recognition accuracy rate of various neural network models to various numbers in the Fashion-MNIST dataset. It can be seen that the recognition accuracy for each label of the model with three ReLU family activation

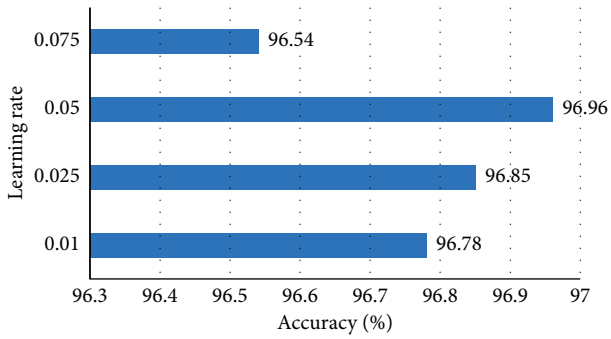

Figure 7: Classification accuracy of the MNIST dataset by the RReLU function N-D^{2}NN model with different learning rates.

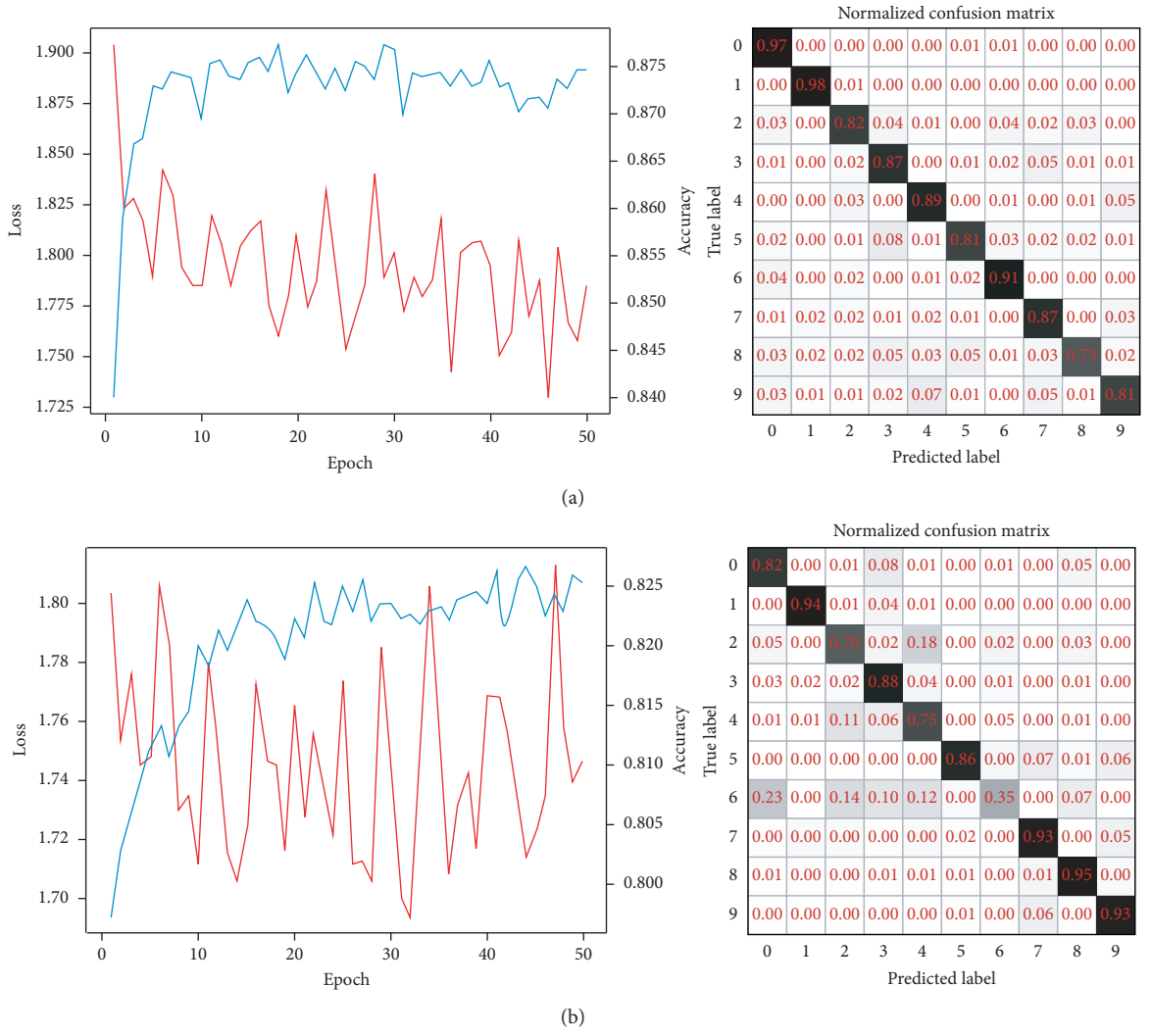

Figure 8: (a) Accuracy rate and confusion matrix of the standard all-optical diffraction deep neural network for the MNIST dataset. (b) Accuracy rate and confusion matrix of the standard all-optical diffraction deep neural network for the Fashion-MNIST dataset.

Table 11: Classification accuracy rates of N-D^{2}NN with different activation functions.

Activation functions	Accuracy (%)	
	MNIST	Fashion-MNIST
Leaky-ReLU	97.76	89.24
PReLU	97.68	89.28
RReLU	97.86	89.28

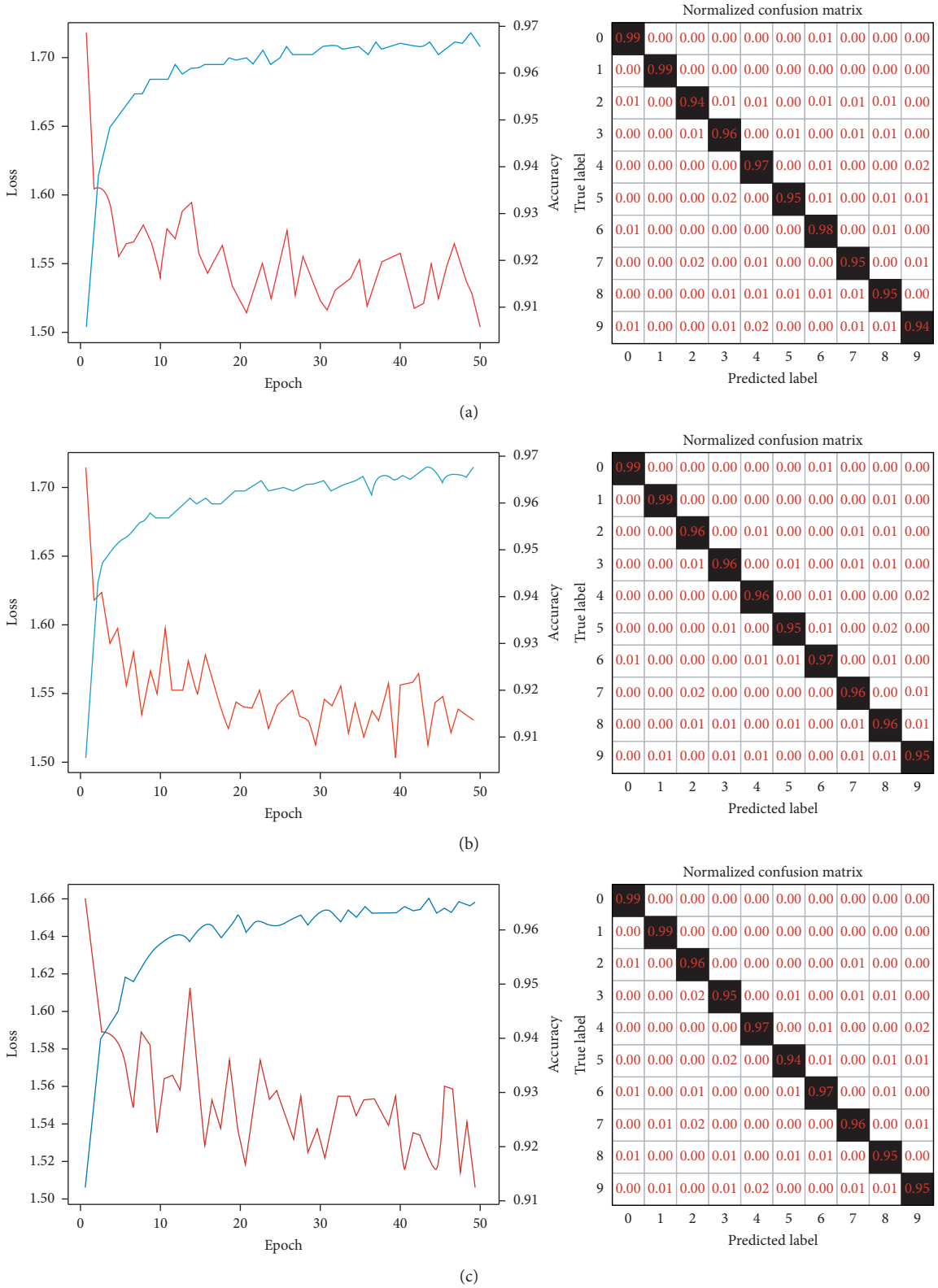

Figure 9: Continued.

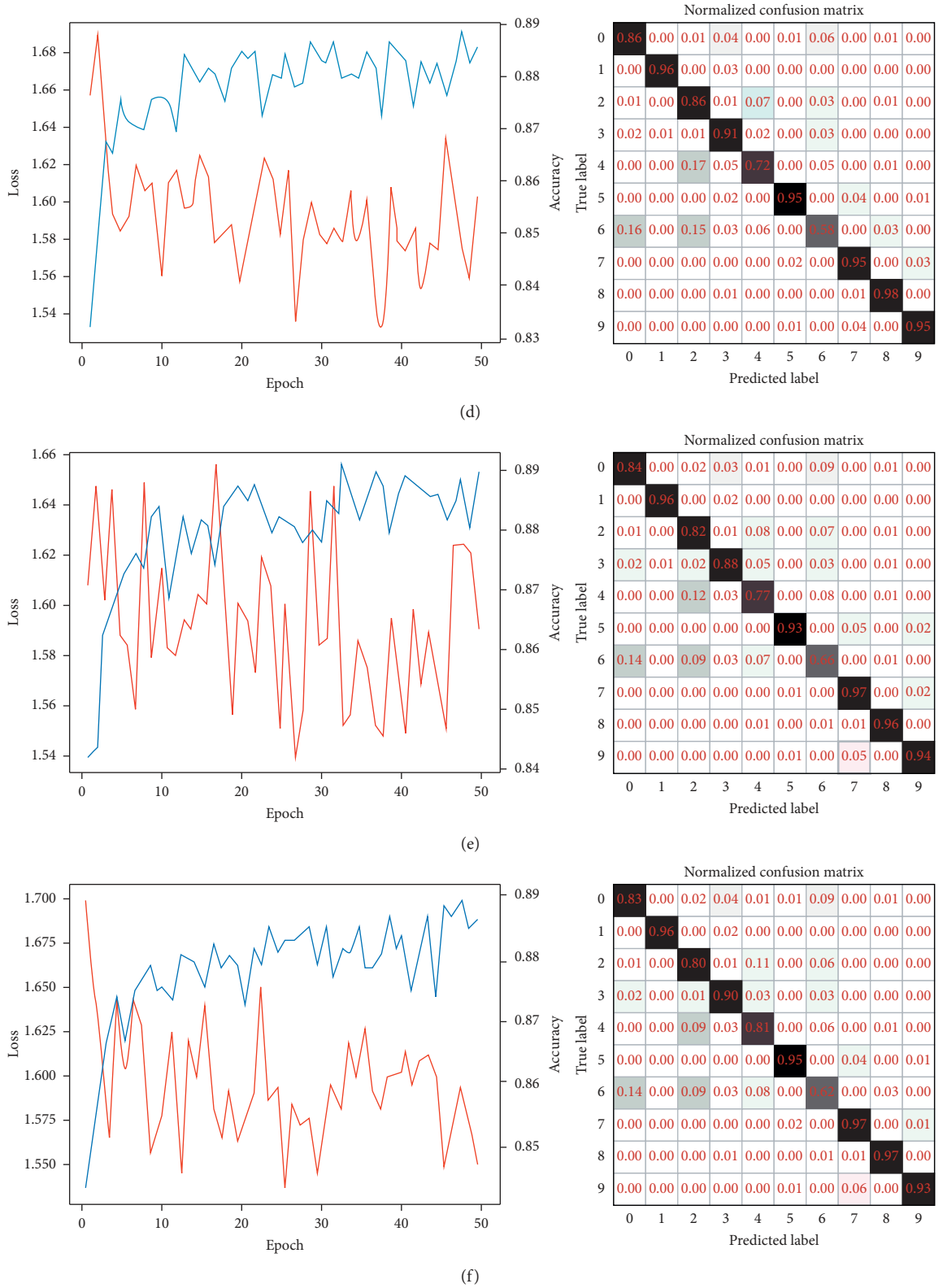

Figure 9: (a) Accuracy rate and confusion matrix of the MNIST dataset by the neural network with the Leaky-ReLU function. (b) Accuracy rate and confusion matrix of the MNIST dataset by the PReLU function neural network. (c) Accuracy rate and confusion matrix of the MNIST dataset by the RReLU function neural network. (d) Accuracy rate and confusion matrix of the Fashion-MNIST dataset by the neural network with the Leaky-ReLU function. (e) Accuracy rate and confusion matrix of the Fashion-MNIST dataset by the PReLU function neural network. (f) Accuracy rate and confusion matrix of the Fashion-MNIST dataset by the RReLU function neural network.

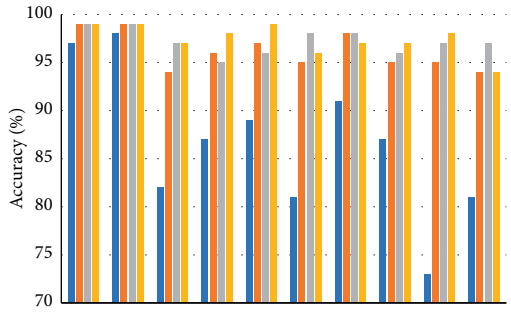

Figure 10: Recognition accuracy of MNIST dataset by N-D^{2}NN.

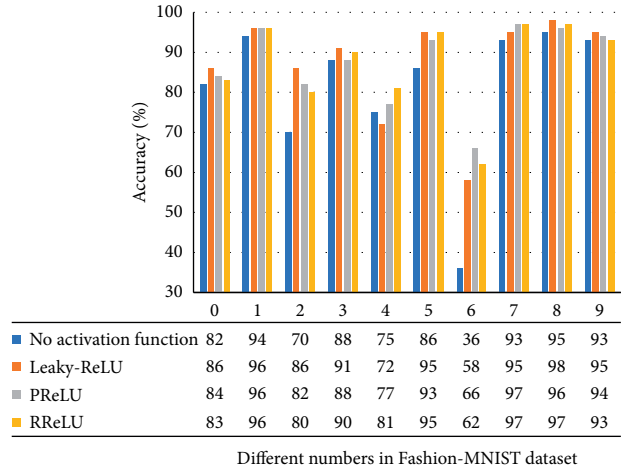

Figure 11: Recognition accuracy of the Fashion-MNIST dataset by N-D^{2}NN.

functions in the Fashion-MNIST dataset is higher than that of the standard model without activation function.

4. Discussion

Nonlinear activation function can improve the representation ability of traditional deep learning. However, in a previous work, optical nonlinearity is not incorporated into deep optical network design, so it is not proved whether the nonlinear effect could improve the representation ability of the N-D^{2}NN framework. In this study, the nonlinear activation function is added to the N-D^{2}NN framework. The represent abilities of the nonlinear N-D^{2}NN framework and the linear N-D^{2}NN framework are analyzed, and it is proved that the nonlinear activation function can improve the representation ability in the N-D^{2}NN framework. The proposed theory can also be extended to any laser with the required wavelength, that is, the diffraction grating suitable for the all-optical D^{2}NN model.

In practice, there are three kinds of methods to realize the nonlinear activated function. The first one is nonlinear material, including crystal, polymer, or semiconductor. Any third-order nonlinear material, which has a strong third-order optical nonlinearity χ (3), can be used to form a nonlinear diffraction layer: glass (As_2S_3, for example, of metal nanoparticles doped glass), polymer (poly two acetylene, for example), organic thin-film, semiconductor (for example, gallium arsenide, silicon, and CdS), and graphene. The second method is saturable absorbent materials, such as semiconductors, quantum dot films, carbon nanotubes, and even graphene films, that can be used as nonlinearity elements for N-D^{2}NN. Recently, a material with the strong optical Kerr effect [43, 44] brings light to the deep diffraction neural network architecture. The third method is that the optical nonlinearity can be introduced into the layers of N-D^{2}NN by using the direct current electrooptical effect. This is an all-optical operation that deviates from the device, and each layer of the diffraction neural network has a direct current field. This electric field can be applied externally to each layer of N-D^{2}NN.

Since, graphene and cadmium sulfide (CdS) have achieved a series of important research results in the field of nonlinear optics. In the following work, the nonlinear saturation absorption coefficient of the above materials will be used to fit the optical limiting effect function, which is used as the activation function in the miniaturized nonlinear diffraction deep neural network. In the simulation state, the classification accuracy of the N-D^{2}NN model for nonlinear optical materials will be verified. One is the method of material coating, that is, a layer of graphene or CdS material is plated on the diffraction grating of germanium material to achieve the physical establishment of the N-D^{2}NN model. Another approach is to directly fabricate diffraction gratings using nonlinear materials such as graphene and CdS.

5. Conclusions

In this study, an N-D^{2}NN structure based on 10.6 μm wavelength nonlinear activation function is proposed based on the optical neural network and complex-valued neural network, and the simulation proves its correctness. The experimental results show that using three ReLU functions, the N-D^{2}NN framework of classification performance is better than that without using a nonlinear activation function N-D^{2}NN framework. This proves the necessity of nonlinear activation function in N-D^{2}NN framework. It can improve recognition accuracy. Comparing with the D^{2}NN model in literature [14, 15], the N-D^{2}NN model using RReLU function can improve the identification accuracy of MNIST dataset by 0.05%. However, there are still two challenges: one is to find the corresponding nonlinear optical materials in the physical model. The other is that there may be a better nonlinear activation function in the N-D^{2}NN framework. These two points are the works that should be completed in the future. In the follow-up study, the neural network model will be further optimized. The nonlinear activation function more suitable for N-D^{2}NN

will be further searched, which provides a theoretical basis for realizing the N-D^2NN physical system of 10.6 μm wavelength.

Data Availability

The raw/processed data required to reproduce these findings cannot be shared at this time as the data also form part of an ongoing study.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This study was supported by a program from the General Project of Science and Technology Plan of Beijing Municipal Education Commission (grant no. KM202011232007), the Programme of Introducing Talents of Discipline to Universities (grant no. D17021), and the Connotation Development Project of Beijing Information Science and Technology (grant no. 2019KYNH204). The authors thank all the participants who have participated in this study.

References

[1] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proceedings of the NIPS*, Curran Associates Inc, January 2012.

[2] K. Cho, B. Van Merrienboer, C. Gulcehre et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, October 2014.

[3] A. Graves, A. R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, Vancouver, Canada, May 2013.

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[5] N. H. Farhat, D. Psaltis, A. Prata, and E. Paek, "Optical implementation of the Hopfield model," *Applied Optics*, vol. 24, no. 10, p. 1469, 1985.

[6] L. Appeltant, M. C. Soriano, d. S. G. Van et al., "Information processing using a single dynamical node as complex system," *Nature Communications*, vol. 2, p. 468, 2011.

[7] A. N. Tait, T. F. D. Lima, E. Zhou et al., "Neuromorphic photonic networks using silicon photonic weight banks," *Scientific Reports*, vol. 7, no. 1, 2017.

[8] A. N. Tait, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Broadcast and weight: an integrated network for scalable photonic spike processing," *Journal of Lightwave Technology*, vol. 32, no. 21, pp. 4029–4041, 2014.

[9] Y. Shen, N. C. Harris, S. Skirlo et al., "Deep learning with coherent nanophotonic circuits," *Nature Photonics*, vol. 11, no. 7, p. 441, 2017.

[10] A. Zanutta, E. Orselli, T. Fäcke, and A. Bianco, "Photopolymeric films with highly tunable refractive index modulation for high precision diffractive optics," *Optical Materials Express*, vol. 6, no. 1, pp. 252–263, 2015.

[11] R. Pashaie and N. H. Farhat, "Optical realization of bio-inspired spiking neurons in the electron trapping material thin film," *Applied Optics*, vol. 46, no. 35, pp. 8411–8418, 2007.

[12] J. Bueno, S. Maktoobi, L. Froehly et al., "Reinforcement learning in a large-scale photonic recurrent neural network," *Optica*, vol. 5, no. 6, pp. 756–760, 2018.

[13] S. Maktoobi, L. Froehly, L. Andreoli et al., "Diffractive coupling for photonic networks: how big can we go?" *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 26, no. 1, pp. 1–8, 2020.

[14] X. Lin, Y. Rivenson, N. T. Yardimci et al., "All-optical machine learning using diffractive deep neural networks," *Science*, vol. 361, no. 6406, pp. 1004–1008, 2018.

[15] D. Mengu, Y. Luo, Y. Rivenson, and A. Ozcan, "Analysis of diffractive optical neural networks and their integration with electronic neural networks," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 26, no. 1, pp. 1–14, 2020.

[16] Y. Luo, D. Mengu, N. T. Yardimci et al., "Design of task-special optical systems using broadband diffractive neural networks," *Light: Science & Applications*, vol. 8, no. 1, pp. 1–14, 2019.

[17] L. Lu, Z. Zeng, L. Zhu et al., "Miniaturized diffraction grating design and processing for deep neural network," *IEEE Photonics Technology Letters*, vol. 31, no. 24, pp. 1952–1955, 2019.

[18] T. L. Clarke, "Generalization of neural networks to the complex plane," in *Proceedings of the 1990 IJCNN International Joint Conference on Neural Networks*, vol. 2, pp. 435–440, San Diego, CA, USA, June 1990.

[19] N. Benvenuto and F. Piazza, "On the complex backpropagation algorithm," *IEEE Transactions on Signal Processing*, vol. 40, no. 4, pp. 967–969, 1992.

[20] G. M. Georgiou and C. Koutsougeras, "Complex domain backpropagation," *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 39, no. 5, pp. 330–334, 1992.

[21] T. Nitta, "A back-propagation algorithm for complex numbered neural networks," in *Proceedings of 1993 International Conference on Neural Networks*, vol. 2, pp. 1649–1652, Nagoya, Japan, October 1993.

[22] I. Aizenberg and C. Moraga, "Multilayer feedforward neural network based on multi-valued neurons (mlmvn) and a backpropagation learning algorithm," *Soft Computing*, vol. 11, no. 2, pp. 169–183, 2007.

[23] N. N. Aizenberg and I. N. Aizenberg, "Cnn based on multi-valued neuron as a model of associative memory for grey scale images," in *CNNA'92 Proceedings Second International Workshop on Cellular Neural Networks and their Applications*, pp. 36–41, Munich, Germany, October 1992.

[24] D. C Park and T. K. Jeong, "Complex-bilinear recurrent neural network for equalization of a digital satellite channel," *IEEE Transactions on Neural Networks*, vol. 13, no. 3, pp. 711–725, 2002.

[25] S. L. Goh, M. Chen, D. H. Popović, K. Aihara, D. Obradovic, and D. P. Mandic, "Complex-valued forecasting of wind profile," *Renewable Energy*, vol. 31, no. 11, pp. 1733–1750, 2006.

[26] Y. Ozbay, "A new approach to detection of ecg arrhythmias: complex discrete wavelet transform based complex-valued artificial neural network," *Journal of Medical Systems*, vol. 33, no. 6, p. 435, 2008.

[27] A. B. Suksmono and A. Hirose, "Adaptive noise reduction of InSAR images based on a complex-valued MRF model and its application t o phase unwrapping problem," *IEEE Transactions*

on Geoscience and Remote Sensing, vol. 40, no. 3, pp. 699–709, 2002.

[28] A. Hirose, "Complex-valued neural networks: the merits and their origins," in *Proceedings of the 2009 International Joint Conference on Neural Networks*, pp. 1237–1244, Atlanta, GA, USA, June 2009.

[29] Z. Zhang, H. Wang, F. Xu, and Y.-Q. Jin, "Complex-valued convolutional neural network and its application in polarimetric sar image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 12, pp. 7177–7188, 2017.

[30] H.-G. Zimmermann, A. Minin, and V. Kusherbaeva, "Comparison of the complex-valued and real-valued neural networks trained with gradient descent and random search algorithms," in *Proceedings of the 19th European Symposium on Artificial Neural Networks*, vol. 18, Bruges, Belgium, April 2011.

[31] B. Xu, N. Wang, T. Chen et al., "Empirical evaluation of rectified activations in convolutional network," 2015, https://arxiv.org/abs/1505.00853.

[32] V. Bianchi, T. Carey, L. Viti et al., "Terahertz saturable absorbers from liquid phase exfoliation of graphite," *Nature Communications*, vol. 8, Article ID 15763, 2017.

[33] J. W. Goodman, *Introduction to Fourier Optics*, Roberts and Company Publishers, Greenwood Village, CO, USA, 2005.

[34] N. Mönning and S. Manandhar, "Evaluation of complex-valued neural networks on real-valued classification tasks," 2018, https://arxiv.org/abs/1811.12351.

[35] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pp. 807–814, Haifa, Israel, June 2010.

[36] Yi Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," 2014, https://arxiv.org/pdf/1505.00853.

[37] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier networks," in *Proceedings of the 14th International Conference on Articial Intelligence and Statistics*, vol. 15, pp. 315–323, JMLR W&CP, Fort Lauderdale, FL, USA, April 2011.

[38] Maas, L. Andrew, Hannun, Y. Awni, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proceedings of the ICML*, vol. 30, Daegu, Korea, November 2013.

[39] K. He, X. Zhang, S. Ren et al., "Delving deep into rectifiers: surpassing human-level performance on ImageNet classification," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, IEEE, Santiago, Chile, December 2015.

[40] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," in *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[41] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms," 2017, https://arxiv.org/abs/1708.07747.

[42] Y. Xiao, H. Qian, and Z. Liu, "Nonlinear metasurface based on giant optical Kerr response of gold quantum wells," *Acs Photonics*, vol. 5, 2018.

[43] D. M. W. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, pp. 37–63, 2011.

[44] X. Yin, T. Feng, Z. Liang, and J. Li, "Artificial Kerr-type medium using metamaterials," *Optics Express*, vol. 20, no. 8, pp. 8543–8550, 2012.