Research Article | Vol. 28, No. 24 / 23 November 2020 / Optics Express 36668
Optics EXPRESS

Integrated photonic neural network based on silicon metalines

SANAZ ZAREI,[1,2] MAHMOOD-REZA MARZBAN,[1,2] AND AMIN KHAVASI[1,*]

[1]Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran
[2]These authors contribute equally to this work
[*]khavasi@sharif.edu

Abstract: An integrated photonic neural network is proposed based on on-chip cascaded one-dimensional (1D) metasurfaces. High-contrast transmitarray metasurfaces, termed as metalines in this paper, are defined sequentially in the silicon-on-insulator substrate with a distance much larger than the operation wavelength. Matrix-vector multiplications can be accomplished in parallel and with low energy consumption due to intrinsic parallelism and low-loss of silicon metalines. The proposed on-chip whole-passive fully-optical meta-neural-network is very compact and works at the speed of light, with very low energy consumption. Various complex functions that are performed by digital neural networks can be implemented by our proposal at the wavelength of 1.55 μm. As an example, the performance of our optical neural network is benchmarked on the prototypical machine learning task of classification of handwritten digits images from the Modified National Institute of Standards and Technology (MNIST) dataset, and an accuracy comparable to the state of the art is achieved.

1. Introduction

Deep neural networks can surpass human intelligence in various applications, including image analysis, medical diagnosis, natural language processing, speech recognition, game playing, etc. and have seen significant growth in recent years. Despite their excellent performance, their complicated structures, especially for performing more complex tasks, requires increasing computing power. This has motivated the search for alternative computing methods that are faster and more energy-efficient. Optical neural networks (ONNs) have been suggested as a low-power, ultra-broad bandwidth, light-speed parallel processing alternative to digitally implemented neural networks [1–17]. Some of the earlier works include ONNs based on free-space diffraction [1–3], chip-integrated photonic platforms using programmable waveguide interferometer meshes [4–7], integrated spiking neural networks using phase-change materials (PCM) [8], convolutional neural networks through diffractive optics [9–10], and artificial neural computing using a nanophotonic neural medium [11].

Metasurfaces are arrays of subwavelength structures capable of shaping the wavefront (phase) of incident electromagnetic waves. Recent research on metasurfaces has paved the way for all-optical signal processing. More complex meta-systems can be realized by cascading multiple layers of metasurfaces. Recently a number of ONN designs based on multi-layered metamaterials and/or metasurfaces have been reported [18–20]. These designs, while performing at the speed of light, are purely passive and do not need power. However, an all-optical integrated on-chip neural network based on metasurfaces is still lacking. Such a design can also drastically reduce the misalignment of layered metasurfaces.

In this paper, we devise a scalable whole-passive fully-optical architecture, based on multi-layered metasurfaces, to realize deep optical neural networks using nano-photonic integrated circuits. In this design, fully-optical matrix multiplications are performed using silicon 1D metasurfaces, named as metalines [21–22]. The presented architecture enables the efficient

#404386 https://doi.org/10.1364/OE.404386
Journal © 2020 Received 3 Aug 2020; revised 5 Oct 2020; accepted 16 Oct 2020; published 18 Nov 2020

Optics EXPRESS

processing and analysis of data in parallel and on-chip, which is fast, low-power and compact. Thus it can find widespread applications in power-critical situations such as mobile devices.

To demonstrate the capability of our ONN, we benchmarked its performance on handwritten digits classification, which achieved an accuracy of 88.8 percent that is comparable to the state of the art. Our 5-layer ONN design has 3920 design parameters and is numerically tested over 10000 test images. For verification of our results, 2.5D variational FDTD solver of Lumerical Mode Solution commercial software is utilized. A reduced-size ONN structure is designed due to computational limitations of Lumerical software package. The reduced design is simulated for 100 handwritten digit images which are selected randomly from the test dataset images that their numerical testing was successful. The simulation results shows 91% matching with analytical analysis results.

The advantages of our proposed ONN to other integrated ONN architectures [4–8] is its simple structure, whole-passive operation, and easy scalability to large number of neurons ($N>1000$). Furthermore, while having the same performance, our ONN has fewer number of neurons (under 4000), comparing to other ONN proposals [3,11,18], which have millions of optical neurons.

2. Proposed ONN architecture

2.1. Optical neural network

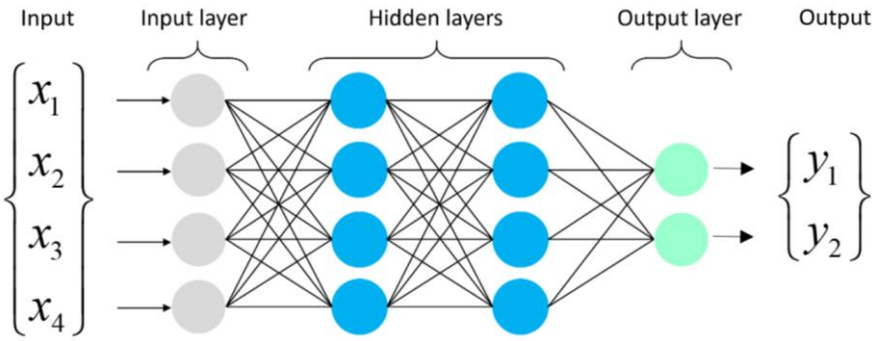
Conventional artificial neural networks are composed of a series of input artificial neurons, connected to one or more hidden layers, and to the output layer at the end [Fig. 1(a)]. Each layer consists of linear matrix-vector multiplications followed by the application of an element-wise nonlinear function, or activation. Conventional ANNs can virtually be implemented in the computer through a software platform. In a classification task, signals travel from the input layer to the output layer, and during this data propagation in the network, each layer of the conventional ANN may perform a different transformation on their inputs. As a result, one can say the neural network performs a series of mathematical operations and map the input data to one of the output classes. The aim of the learning procedure is to produce the desired output for each input. A cost function is chosen to quantify the difference between the target output and the output predicted by the network. This cost function is chosen to eliminate incorrect deductions, and in the learning phase, it is tried to minimize the cost function by finding appropriate weight matrices.

A schematic view of our proposed ONN design is presented in Fig. 1(b). The task which is an image in our work, is preprocessed and converted to a vector. The preprocessed signals are then encoded to the amplitude of the optical pulses which propagates through our multi-layer photonic integrated neural network. Each layer of the ONN consists of a metaline (a line of meta-atoms) created on a *SOI* platform, which its design details are described in the next subsection. The ONN make use of wave propagation and diffraction to perform appropriate mathematical operations between the input and output of the device. Since our proposed ONN is a linear diffractive optical network, all physical phenomena like wave propagation and diffraction that happen between the input and output of the device can be squeezed into a single matrix operation [2], which can be decomposed to matrix multiplications. Multiple silicon metaline layers are used to implement the fully-optical matrix multiplications.

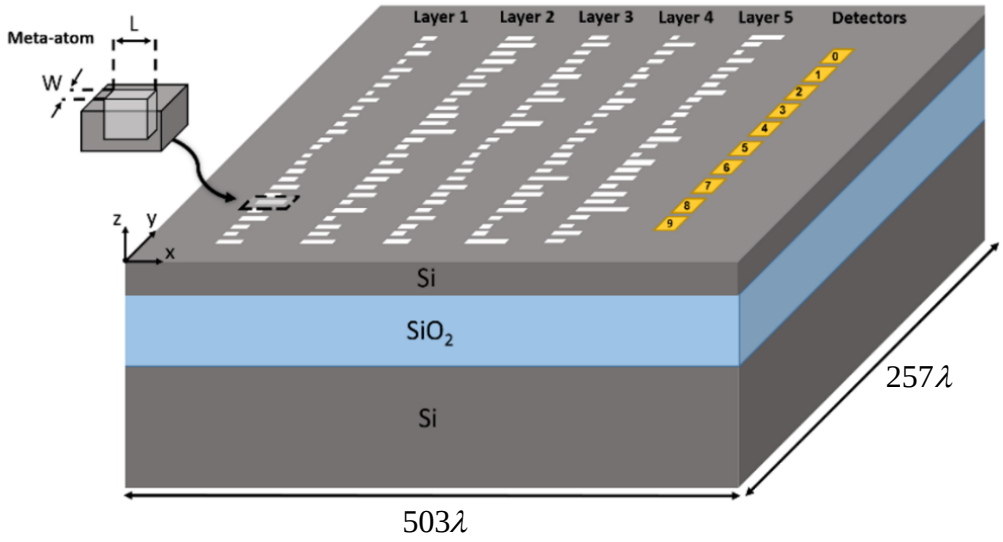
2.2. Metalines

The optical architecture used in this work composed of 1D high-contrast transmitarray metasurfaces [21], termed as metalines, that are 1D etched rectangle slot arrays in silicon-on-insulator (SOI) substrate. The thickness of the silicon membrane of SOI is 250 nm. The SiO_2 insulator layer is 2 μm thick. Each unit cell of the metalines, i.e., meta-atom, is geometrically parameterized by the lattice constant of the metalines (a), length (L), and width (w) of the slots. Here, we fix the lattice constant of the metalines to approximately one-third of the wavelength (500 nm) to keep

Optics EXPRESS



(a)



(b)

Fig. 1. Schematic of (a) a conventional artificial neural network, (b) the proposed optical neural network in this paper. The ONN constitutes five metaline layers, each of which having several meta-atoms.

the structure subwavelength. If the length and width of slots within meta-atoms are swept, the transmission amplitude and phase of the metalines for TE-polarized guided waves are as shown in Figs. 2(a) and 2(b). The operation wavelength is set to be 1.55 μm. By fixing the width of the slots to 140 nm and by changing the length of the slots from 0.05 μm to 2.5μm, free control of the propagation phase within the $0-$ to -2π range with transmission amplitude higher than 0.96 can be achieved. These are depicted in Figs. 2(c) and 2(d). The results presented in Fig. 2 are obtained using the commercial software package Lumerical FDTD.

Our proposed ONN is formed by several metalines, where each meta-atom on a given metaline is represented by a complex transmission coefficient ($T^l = t^l \exp(j\phi^l)$). As can be inferred from Fig. 2(d), for the chosen geometrical parameters, the transmission amplitude is very near to 1. Thus it can be assumed that the transmission loss is negligible for all the meta-atoms, and the output wave of each meta-atom can be determined only by the propagation phase shift of its input wave ($T^l = \exp(j\phi^l)$). In each metaline, by carefully choosing the slot lengths for meta-atoms at

Optics EXPRESS

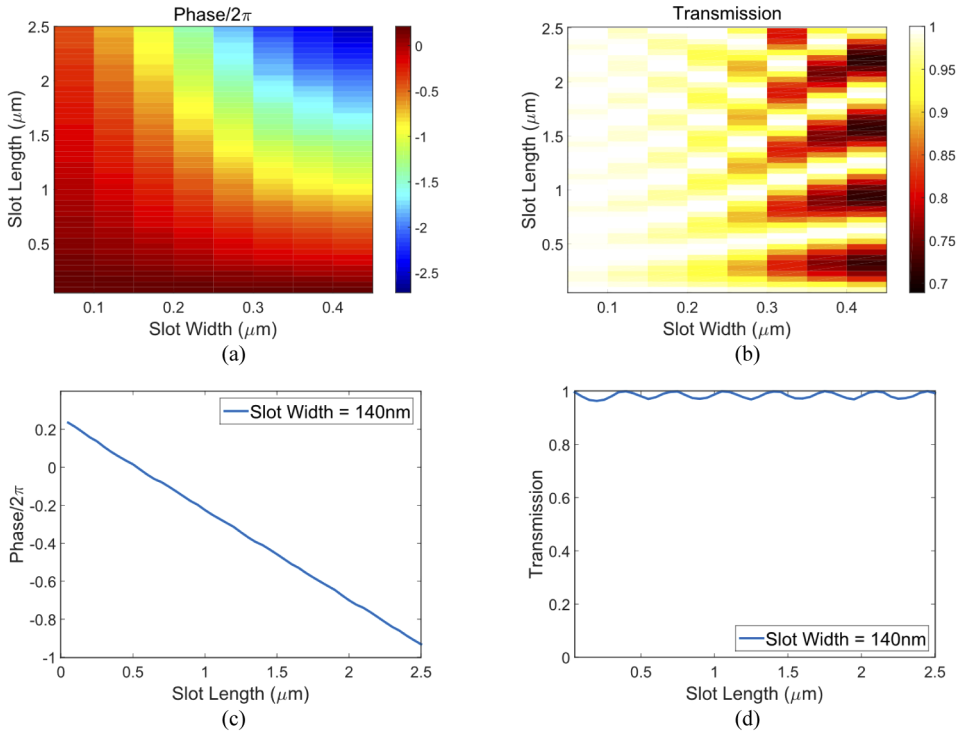


Fig. 2. Phase (a) and transmission amplitude (b) of a meta-atom for TE-polarized guided wave versus slot length and width. Phase (c) and transmission amplitude (d) versus slot length, when fixing the slot width to 140 *nm*. The operation wavelength is 1.55 μm.

different positions, an arbitrary phase profile can be achieved. The other assumption made for our metaline-system is that the output phase associated with a meta-atom depends only on its own geometrical parameters, and cannot be affected by its neighbors. This assumption can be explained as the optimal metaline is locally periodic over most regions.

3. Modeling

For the electromagnetic modeling of our ONN, we follow a very similar approach to the one presented in [18]. In this work, input images are pixelated, and each pixelated image, consisting of $N \times N$ pixels, is converted to a vector and then is encoded into the amplitude of the input electric field to ONN. Thus, the input electric field (E^{in}) is a $N^2 \times 1 -$ dimensional vector. The input line can be placed at the distance d of the first layer of the multi-layered metalines.

Our design problem is assumed to be two-dimensional (2D). The input field propagates through multiple layers of metalines. At the place of m'th metaline, the incident electric field sees a line of spatially-varying phase shifts determined by the length of the slots in the meta-atoms, which are denoted by the vector w^m. Each metaline has N^2 meta-atoms, so that w^m is also a $N^2 \times 1 -$ dimensional vector. At the output line of the system, the output electric field (E^{out}) is the function of the parameters $\{w^1, w^2, \ldots, w^M\}$ of the M metalines. E^{out} can be calculated following a series of matrix-vector multiplications as below (Fig. 3):

$$E^{out} = \left(\prod_{m=1}^{M} F^{+} P^{m} F \Phi^{w^m} \right) E^{in} \tag{1}$$

Optics EXPRESS

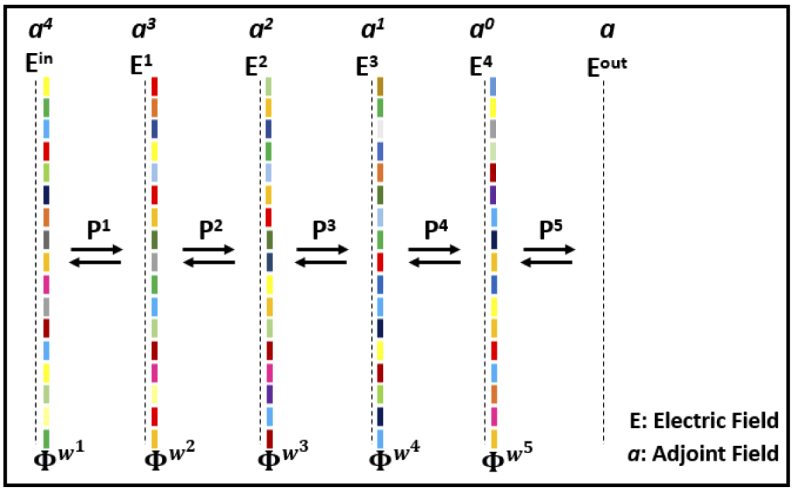


Fig. 3. The forward propagation model utilized in this paper. For error back-propagation, the adjoint field is back propagated through the system.

In which F and F^{+} are the discrete Fourier transform and its inverse. P^{m} is a diagonal matrix that takes into account the plane wave propagation from one metaline layer to the next layer (see [18]), and Φ^{w^m} is a diagonal matrix, containing complex exponentials associated with the phase shifts induced by the meta-atoms on the m'th layer. We can express $\Phi^{w^m} = \exp(j\phi^{w^m})$, in which ϕ^{w^m} is another diagonal matrix containing the phase shifts associated with w^m.

In the mathematical model expressed above, each meta-atom on a single layer introduces a phase modulation on the input electric field at that meta-atom. The produced electric field at the output of the corresponding metaline is decomposed into a superposition of plane-waves using the discretized Fourier transform [18]. Each plane-wave is multiplied by an appropriate phase factor accounting for the phase accumulated as the wave propagates to the next layer. An inverse Fourier transform operation is performed to determine the resulting electric field at the input of the next metaline layer [18].

Once E^{out} is computed, the output intensity at the s'th sample position on the output line can be calculated as $I_s = E_s^{out*} E_s^{out}$. The cost function in our optimization is defined as the squared errors between the desired set of output intensity distributions and a realized set of intensity distributions. In image classification tasks, cost function should be computed for each input image and then summed over all the input images (K):

$$C = \sum_{k=1}^{K} \sum_{s=1}^{N^2} (I_s^k - I_s^{des,k})^2 \tag{2}$$

The cost function must be iteratively minimized, adjusting the design parameters $\{w^1, w^2, \ldots, w^M\}$. The mini-batch gradient descent algorithm [23], is used for this iterative optimization. For Back-propagating the errors in the network to update these learnable parameters, calculation of the gradients for all design parameters is needed. Therefore an adjoint gradient method is utilized. To compute dC/dw^m, we use the following chain rule:

$$\frac{dC}{dw^m} = \frac{d\vec{\phi}^{w^m}}{dw^m} \otimes \frac{dC}{d\vec{\phi}^{w^m}} \tag{3}$$

Optics EXPRESS

where $\vec{\phi}^{w^m}$ denotes an $N^2 \times 1-$ dimensional vector containing the diagonal entries of the matrix ϕ^{w^m} and symbol $\otimes$ denotes element-wise multiplications. The term $d\vec{\phi}^{w^m}/dw^m$ can be easily calculated using Fig. 2(c). To calculate the second term in Eq. (3), the chain rule is utilized again:

$$\frac{dC}{d{\phi_{s'}}^{w^m}} = \sum_{s=1}^{N^2} \left(\frac{dC}{dE_s^{out}}\right)\left(\frac{dE_s^{out}}{d\phi_{s'}^{w^m}}\right) \tag{4}$$

where ${\phi_{s'}}^{w^m}$ is the phase shift at a single spatial location s' on the m'th metaline. We make use of Eqs. (1) and (2) to derive the derivatives in Eq. (4), which its details can be found in Ref. [18]. The finally obtained expression $dC/d\vec{\phi}^{w^m}$ is as:

$$\frac{dC}{d\vec{\phi}^{w^m}} = -4R\left\{i\left(E^{m-1,k+}\right)^T \otimes \left(\prod_{m'=0}^{M-m} \Phi^{w^{M-m'}} + F^+ P^{M-m'+} F\right) a\right\} \tag{5}$$

where $R\{\}$ denotes the real part of the expression in the brackets, $a^k = E^{out,k} \otimes (I^k - I^{des,k})$ is the adjoint field and $E^{m-1,k}$ is the intermediate electric field incident upon the m'th metaline and $^+$ denotes the adjoint of a matrix. In this equation, in order to calculate the gradient $dC/d\vec{\phi}^{w^m}$, the adjoint field is propagated backward through the optical system. By performing element-wise multiplication of the intermediate Foreward-propagated fields with backward-propagated adjoint fields at each metaline, $dC/d\vec{\phi}^{w^m}$ is evaluated. This propagation technique reduces the computational requirements of a single gradient descent iteration to just one forward simulation and one adjoint backward simulation.

4. Design of ONN for handwritten digit classification

To demonstrate the capability of our ONN to perform desired tasks, we train our network as a digit classifier, where ten handwritten digits, from 0 to 9, are chosen for recognition. The training is accomplished by the computer to advance the design. Once the design is finalized, the inference/prediction is fully optical.

The optimized ONN design consists of 5 metaline layers, with each metaline containing 784 meta-atoms (392 μm length). The distance between two successive layers is 150 μm. After light exits the final (fifth) metaline, it propagates 180 μm until it reaches the output line of the network with ten detector regions arranged in a linear configuration. A specific digit is assigned to each detector. The length of each detector is 7 μm, and the distance between two neighboring detectors is 7 μm.

Training is performed using 60,000 (training and validation) images from MNIST (Modified National Institute of Standards and Technology) handwritten digit database [24], where a subset of 4,000 images is randomly selected as the training dataset. Each pixelated image consisting of 28×28 pixels is normalized, converted to a vector, and then encoded as the intensity of input light. For each image, the desired intensity distribution $I_s^{des,k}$ is defined at the output layer of the ONN as a Gaussian centered over the appropriate output detector with variance parameter $\sigma^2 = 6.25\ \mu m^2$.

The lengths of the slots in meta-atoms are chosen as the learnable parameters during the learning process. In practice, a slot with a length smaller than 60 nm would be considered challenging for the fabrication process, so some restrictions are set in the learning procedure to make sure that each slot length is larger than 60 nm. The optimization problem involves 3920 design variables (784 meta-atoms per metaline, and five metalines). For a specific digit, when the output signal is accurately distributed such that the total intensity incident upon the

Optics EXPRESS

expected detector corresponding to that digit has the most significant value comparing to the other detectors, the classification can be considered successful.

Following the mathematical framework presented in section 3, the phase of meta-atoms (and correspondingly the length of the slots) are adjusted in search of minimum cost function corresponding to squared errors between the desired set of output intensity distributions and realized set of intensity distributions at a training iteration. The cost function is iteratively minimized using mini-batch gradient descent algorithm [23,25]. Each batch consists of 64 different input images.

Figure 4 shows the cost values for the training set and the accuracy values for the test set for each Epoch during the learning procedure.

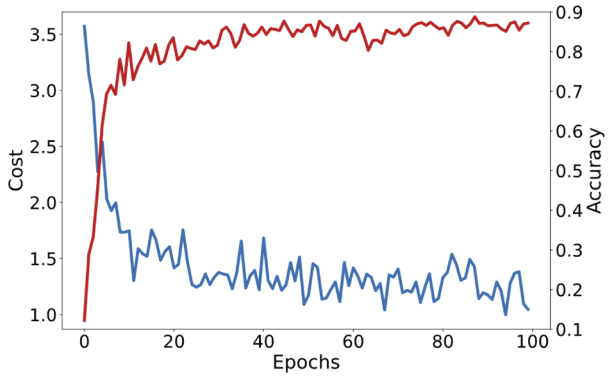


Fig. 4. The cost value on the training set (blue line) and accuracy on the test set (red line) for the optimized 5-layer ONN at each epoch during the learning process.

After each Epoch, the design is tested using 1,000 handwritten digits, which were randomly selected from 10,000 images contained in the MNIST test database. The classification accuracy of the optimized five-layer ONN is 88.8% over the test data set after 88 Epochs, and the cost value is less than 1.5.

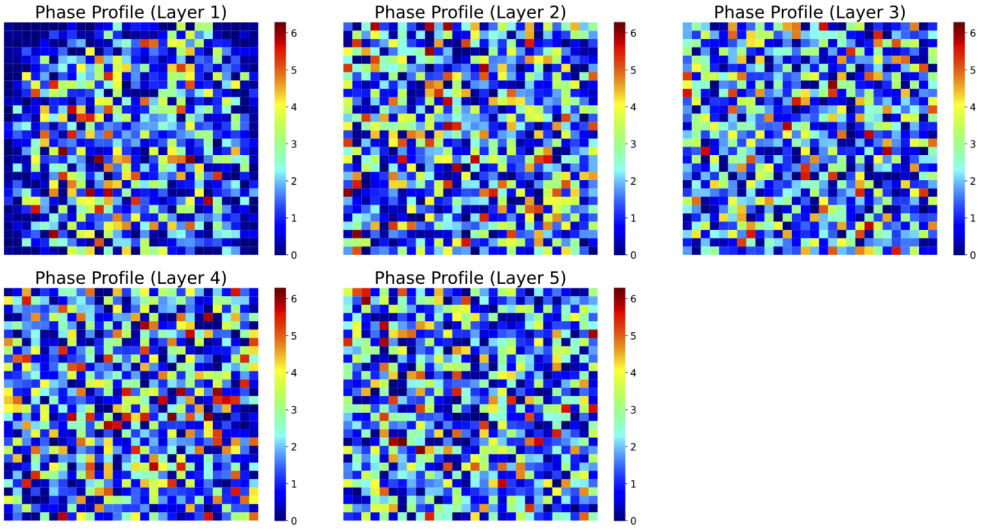


Fig. 5. The phase profile of each ONN metaline layer after training. The linear phase profile is converted to a 28×28 pixelated image.

Optics EXPRESS

The phase profile of each ONN metaline-layer containing 784 phase elements ranging between 0 and 2π, is depicted in Fig. 5. The linear phase profile for each layer is reshaped to a 28×28 pixelated image.

To illustrate the overall performance of our ONN design, three representative input images are chosen from the MNIST test data set [Fig. 6(a)], and the resulting output intensity distribution of the network for these digits are depicted in Fig. 6(b). As it is clear from the pictures, for each digit, a different detector achieves the maximum intensity which facilitate classifying the digits correctly.

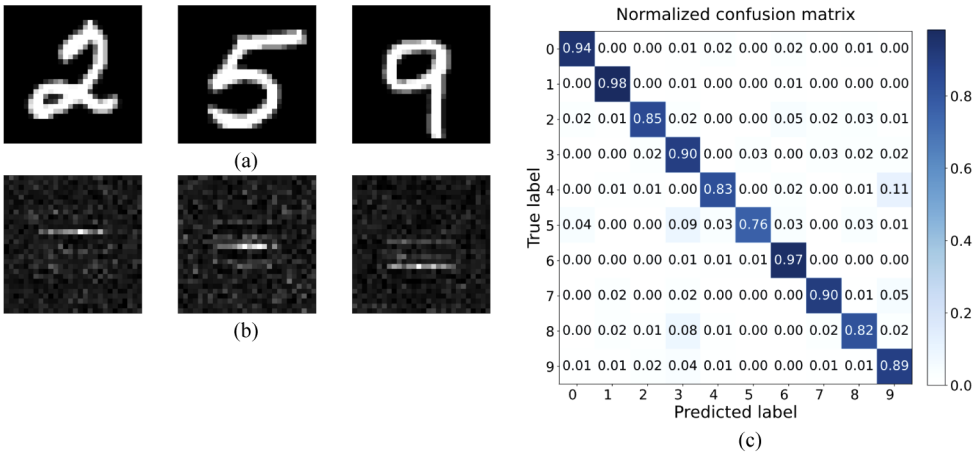


Fig. 6. (a) Input digit, (b) the intensity distribution at the output of optimized five-layer ONN for each corresponding input image (the output vector of 784 elements reshaped to an image of 28×28 pixels), (c) normalized confusion matrix of the ONN over 1000 handwritten digit images from MNIST testing dataset.

Figure 6(c) shows the normalized confusion matrix for the optimized 5-layer ONN over 1000 handwritten digit images of the test data set. The matrix main diagonal shows the probability of correct prediction for each input digit image. Digit 5 is the most challenging digit for classification. Also as is indicated in Fig. 6(c), the largest errors occur while distinguishing between "4" and "9" and "5" and "3" as a result of the similarities in the shapes of these digits.

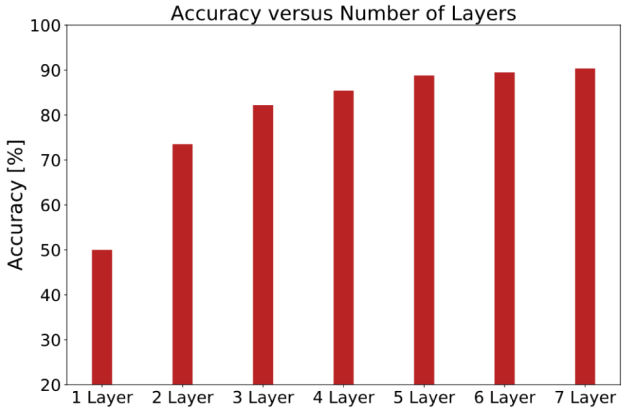


Fig. 7. The accuracy of the proposed ONN architecture on the MNIST test data set, as a function of the number of metaline layers.

Optics EXPRESS

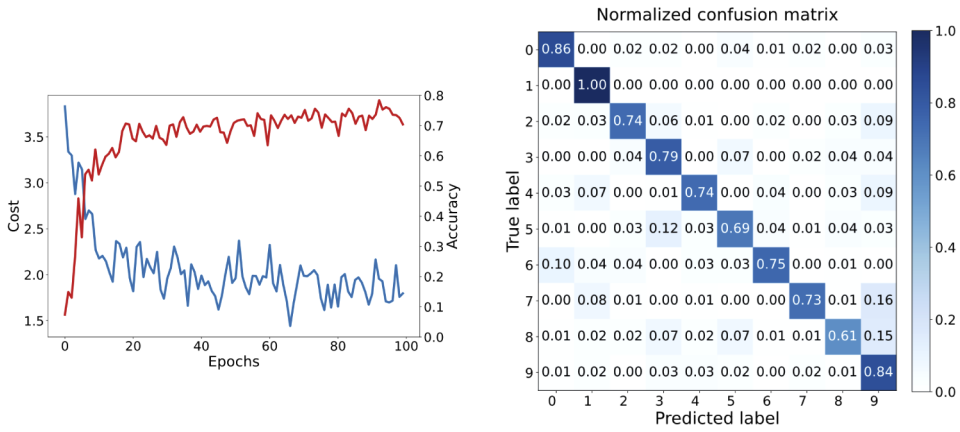
To observe the dependence of classification accuracy on the layer number, ONNs with 1, 2, . . . , 7 layers are also designed and optimized. As shown in Fig. 7, the accuracy keeps increasing from 50.7% for one-layer ONN to 90% for 7− layer structure. Also, it can be inferred from Fig. 7 that the increase rate of accuracy with respect to layer number is much slower for ONNs with more than 3 metaline layers. Thus, the optimal design of the ONN may be chosen based on the minimum requirements of classification accuracy and power efficiency.

5. Design verification

Like all metasurface structures, our proposed ONN inherently has multiscale nature. The whole structure has macroscale dimensions while it consists of nanoscale individual meta-atoms. Modeling such structures that have nanoscale and macroscale length-scales simultaneously encounters many computational restrictions. In this section, to ensure the design procedure of our ONN and to analyze its operation, a reduced structure is designed using the electromagnetic model described in section 3 and the obtained results are verified by 2.5D variational FDTD solver of Lumerical Mode Solution. Reduced structure refers to a structure with reduced dimensions, that is chosen based on the computational limitations. The characteristics of the reduced structure are summarized in Table 1.

Table 1. The reduced structure's characteristics.

Number of layers	3
Number of meta-atoms per layer	196
Number of design variables	588
Length of each metaline-layer	98 μm
Distance between two neighboring layers	20 μm
Distance between the last layer and output layer	60 μm
Total device size	$64\lambda \times 78\lambda$
Number of detectors	10
Detectors arrangement	Linear



(a) (b)

Fig. 8. (a) The cost value on the training set (blue line) and accuracy on the test set (red line) of the reduced ONN at each epoch during the learning procedure, (b) the confusion matrix of the optimized structure.

Optics EXPRESS

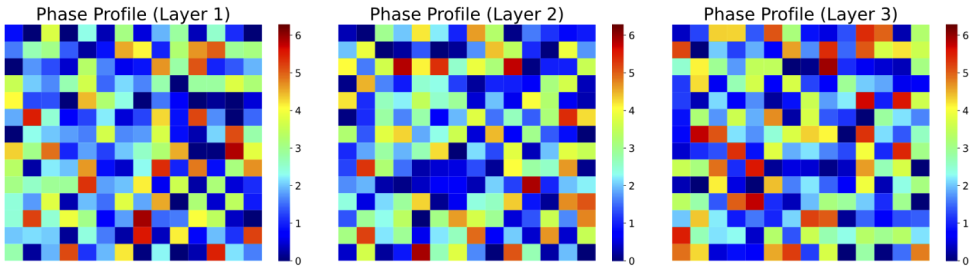


Fig. 9. The optimized phase profile for the first, second and third metaline layers.

Like the previous section, the reduced ONN is trained as a digit classifier for 10 handwritten digits 0 to 9. The training process is the same as before, except the input images which are down-sampled to 14×14 pixels to match the dimensions of the reduced ONN. Also for each image, the desired intensity distribution at the output layer is a Gaussian centered over the appropriate output detector with variance parameter $\sigma^2 = 4\ \mu m^2$. Figure 8(a) presents the cost values for the training set and the accuracy for the test set for each Epoch during the learning procedure. The classification accuracy reaches 78.4% in 93 epochs for the reduced ONN structure. The confusion matrix of the optimized ONN is given in Fig. 8(b). As can be inferred from this figure, digits 8 and 5 are the most challenging numbers for prediction, respectively. Figure 9 shows the phase profile of the optimized metaline layers.

To verify the obtained classification results, the optimized ONN is implemented in Lumerical Mode Solution. 100 handwritten digit images, ten different images per digit, are selected from the 78.4% of the test data set images that their numerical testing was successful. The input images are normalized and down-sampled to 14×14 pixels, and then they are converted to a vector. This vector is encoded as the intensity of input light in Lumerical Mode slab plane wave sources. Ten $y-z$ frequency domain field and power monitors are inserted at the output plane of the structure representing the ten detectors that are assigned to each digit. The distance between two sequential monitors is $2\mu m$, and the width of each monitor is 5 μm. For each input image, the simulation is run over the entire ONN structure.

The simulation results indicates that for 91 out of 100 cases, the 2.5D variatioanl FDTD solver of Lumerical Mode Solution has similar predictions as the electromagnetic model of section 3, which means 91% match between the two results. The mismatch between these two results could be explained by the fact that we did not include some phenomena such as mutual coupling between adjacent meta-atoms and reflection between layers to simplify our electromagnetic model. As a result, adding more layers to the ONN could increase the mismatch between the results of the electromagnetic model and the 2.5D variatioanl FDTD solver of Lumerical Mode Solution. However, as presented in Fig. 10 in most of the cases this mismatch would not cause an impact error and the simulation results are the same as the electromagnetic results. The confusion matrix of the reduced ONN simulated by Lumerical Mode Solution for the chosen 100 images is represented in Fig. 11. As can be seen, the reduced network has the greatest difficulty predicting 5 and 8. This is in good agreement with numerically predicted results.

Figure 12 shows the $x-y$ electric field distribution of the simulated ONN for three representative handwritten digit examples. The detector that receives the maximum light intensity comparing to the other detectors is considered as the network prediction for that specific input digit.

Optics EXPRESS

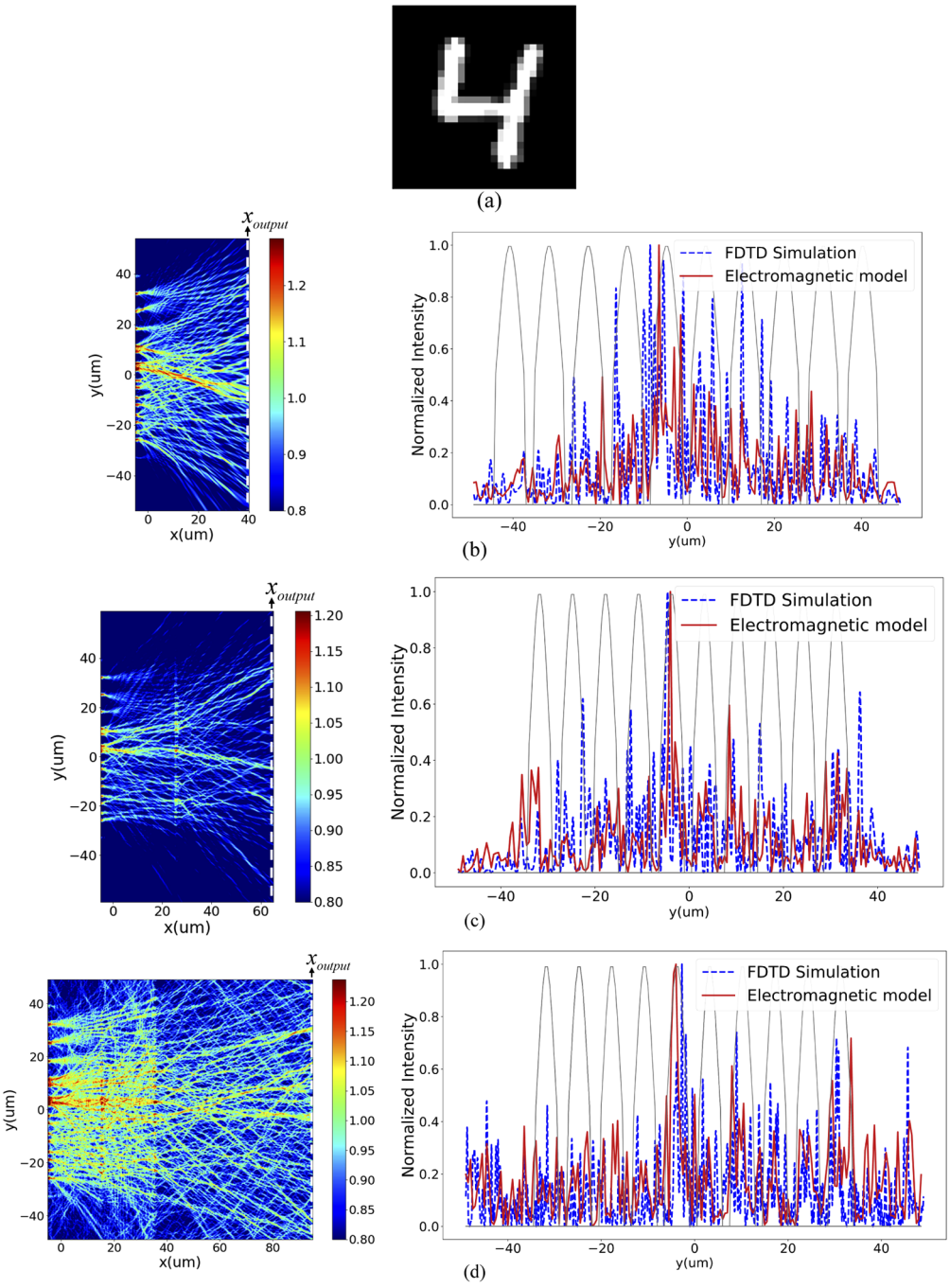


Fig. 10. (a) The input image to the designed ONN, $x - y$ electric field distribution of the simulated ONN based on 2.5D variational FDTD simulations for the input digit shown in (a), and the output field intensity obtained by electromagnetic modeling described in section 3 (red solid lines) and Lumerical Mode Solution (blue dotted lines) at $x = x_{output}$ for (b) 1-Layer, (c) 2-Layer, (d) 3-Layer network. As is seen, the highest intensity corresponds to digit 4. The gray lines indicate the ideal response of the network for each digit.

Optics EXPRESS

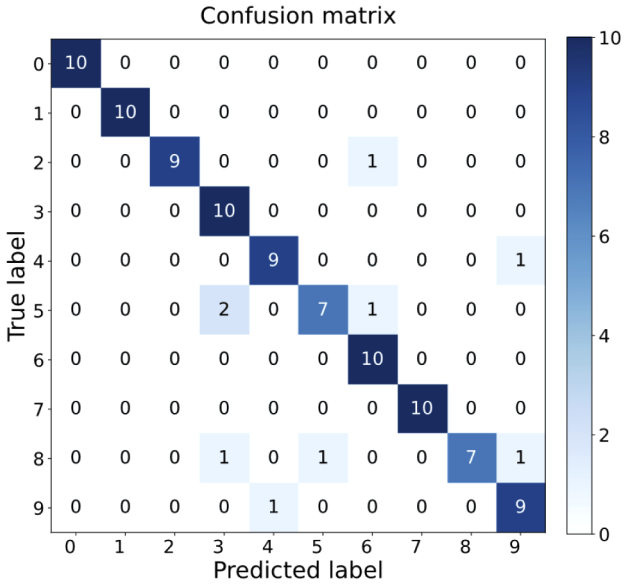


Fig. 11. Confusion matrix for the chosen 100 images from the test data set, generated based on the results of Lumerical Mode Solution simulations.

Optics EXPRESS

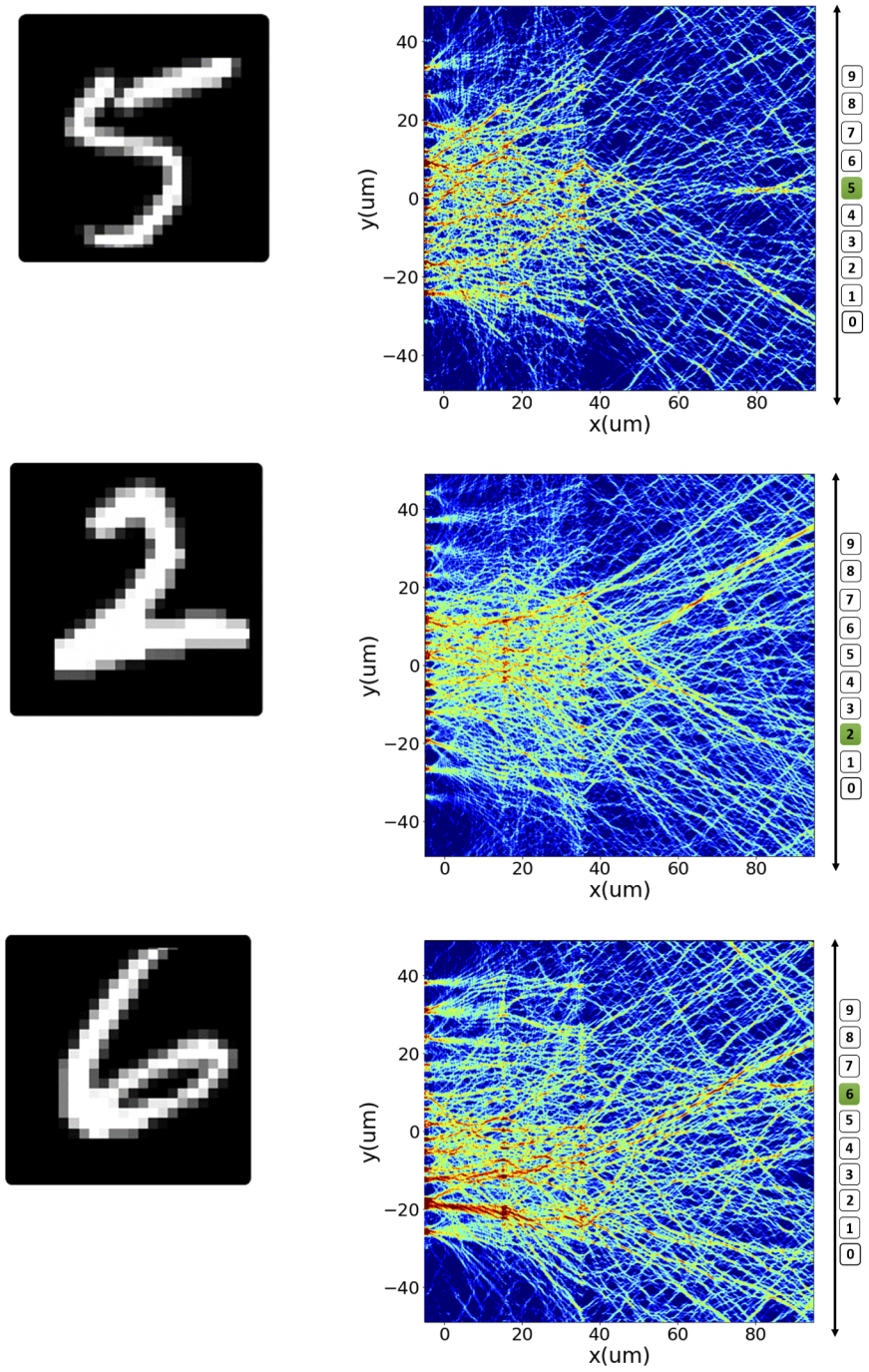


Fig. 12. The $x - y$ electric field distribution of the simulated ONN showing its prediction performance for three representative digits.

Optics EXPRESS

6. Discussion

6.1. Power consumption

Once the optimized design is finalized and fabricated on an SOI substrate, the computation through the proposed optical neural network is fully optical using a light source and optical diffraction through passive components (which can perform computations on the optical signals without additional energy input). Thus, the power consumption of our ONN is only due to the optical source which supplies the input vector (E^{in}).

6.2. Loss

There are some losses in the network, which are mainly due to transmission loss of meta-atoms and the propagation loss through the SOI substrate. The maximum transmission loss imposed by each meta-atom (computed by Lumerical FDTD) is approximately $0.16dB$ at the wavelength of 1.55 μm, and due to parallel interaction between electric field and meta-atoms on a metaline layer, it can be concluded that the maximum transmission loss of each layer is $0.16dB$. Therefore, the total loss can be estimated as:

$$Loss = Transmission\ Loss + \ \mathrm{Pr}\,opagation\ Loss = L \times 0.16dB + \mathrm{Pr}\,opagation\ Loss \quad (6)$$

where L is the number of layers. The other cause of loss is the propagation loss in the network, which can be neglected due to negligible loss of silicon at 1.55 μm.

However in practice, the loss in the system is higher than the above-mentioned value. The reason is that the local periodic approximation is not absolutely true over all regions. This means that the output phase associated with a meta-atom not only depends on its own geometrical parameters, but also on the geometrical parameters of its neighbors. This issue can be more problematic for larger networks. To decrease such losses, some strategies are suggested in [18], including specifying constant or near constant meta-atoms geometries over a supercell of multiple meta-atoms. However, this leads to larger networks.

One advantage of our on-chip ONN compared with other ONNs based on multi-layered two-dimensional metasurfaces [18–20] and/or diffractive layers [1–3] is the easy alignment of the on-chip metalines with precise spacings, due to the existing advanced lithography techniques. This will dramatically reduce the misalignment loss that results in performance degradation. Backer [18] reported about 11% reduction in handwritten digit classification accuracy of his designed ONN, caused by misalignment of layers.

6.3. Latency

For our ONN, the latency is defined by the time between supplying an input vector, E^{in}, and detecting its corresponding output vector, E^{out} (computing an inference result) [6]. This is the travel time for an optical input through all layers. If we assume the propagation distance between the layers is D_p and the propagation distance between the last layer and output layer is D_f and the maximum propagation distance through each layer is D_m, then we have:

$$Latency = L \times D_m \times v_{g1}^{-1} + ((L-1) \times D_p + D_f) \times v_{g2}^{-1} \quad (7)$$

in which $v_{g1} = \frac{c_0}{n_{eff1}}$ and $v_{g2} = \frac{c_0}{n_{eff2}}$ are the speed of light in the metalines and silicon device layer, respectively. For our designed five-layer optical neural network, the latency is approximately $7.78ps$, which is orders of magnitude lower than the typical latency associated with GPUs [6]. Also, this value is far less than the values reported for nanophotonic circuits based on Mach-Zehnder interferometers [6] and the convolutional neural networks based on metasurface optics [10].

Optics EXPRESS

6.4. Computational speed

The speed of ONN is defined as the number of input vectors, E^{in}, that can be processed per unit time. Due to fast response of our ONN (low latency), state of the art high-speed photodetectors are the limiting factor in the speed of our architecture. By assuming $20GHz$ photodetection rate, our ONN with N meta-atoms per layer (N^2 matrix multiplications) can effectively perform $N^2 \times L \times 2 \times 10^{10} MAC/\text{sec}$($MAC$ is the multiply-accumulate operation, which consists of a single multiplication and a single addition.). This means that each layer of our ONN with $N = 784$ will scale the performance of the network to $1.2 \times 10^{16} MAC/\text{sec}$ per layer. This is four orders of magnitude greater than the peak performance obtainable with modern GPUs, which typically have performance on the order of 10^{12} floating point operations/sec (FLOPs) [6].

6.5. Footprint

The total footprint of our five-layer ONN can be computed as:

$$A = N \times a \times (L \times D_m + (L-1) \times D_p + D_f) \approx 400\mu m \times 800\mu m \tag{8}$$

where a is the meta-atom pitch which is 0.5 μm in our work. This footprint is smaller than the previously proposed on-chip optical neural networks [4–8], diffraction-based [1–3] and metasurface-based ONNs [18,20].

For example, in the on-chip photonic neural networks based on Mach-Zehnder interferometers [4–7], the large footprint of directional couplers and phase modulators cause each neuron to have large footprint of 100×60 μm^2. In the on-chip optical neurosynaptic network presented in [8], a neuron with 4 inputs has footprint of approximately 2.5 $mm \times 250$ μm. These make scaling to large number of neurons ($N \geq 1000$) very challenging [13].

6.6. Accuracy

The presented five-layer optical neural network with 3920 neurons have a handwritten digit classification accuracy of 88%, which is higher than the reported accuracy of 84% for a 10-layer metasurface-based ONN at the wavelength of 560 nm [18] and comparable to the accuracy of 89% reported in [19] for a five-layer metasurface-based ONN at the wavelength of 700 nm. In [3], for a diffractive network with five diffraction layers, a higher classification accuracy of 91.75% at the wavelength of 0.75 mm is declared, however, by utilizing the total neuron numbers of 0.2 $million$.

The maximum handwritten digit classification accuracy offered by nanophotonic circuits presented in [6] was 85.83% for linear ONN and by addition of nonlinear activation functions to the circuits, the accuracy increases to 93% . In our work, we didn't use nonlinear activation. However, use of optical nonlinearities as the nonlinear activation function can possibly improve the overall performance.

6.7. Applications

Despite the advantages offered by our on-chip meta-neural-network architecture, like its simple structure, compact size, operation without an alignment step, and good tolerance to fabrication-related geometric distortions [because of its large critical dimensions (>100 nm)] [21], it is limited to process one-dimensional (1D) data. Although this can still have wide range of applications including 1D signal processing, speech recognition and processing, natural language processing, optimization, real-time control of multi-sensor, multi-actuator systems, etc., it has limited application scenarios in image processing problems, as a result of encoding the input images to one-dimensional arrays. However, in [6,26] they benefit from some image pre-processing to overcome such limitations of their on-chip integrated design. Besides, owing to low power consumption and light-speed processing of our architecture, it can be used for analog computing with high performance.

Optics EXPRESS

7. Conclusion

In conclusion, a whole-passive fully-optical architecture for realizing deep neural networks, based on on-chip multi-layered metalines, is presented in this paper. This integrated photonic neural network is capable of performing various complicated tasks at the speed of light, and has a very few power consumption due to optical transportation of data. Also, it is very compact with a small footprint of $400 \times 800\ \mu m^2$. Furthermore, it offers advantages like simple structure, whole-passive operation and easy scalability to large number of neurons, comparing with other integrated optical neural networks; and much fewer numbers of neurons at the same performance, comparing with other diffraction-based and metasurface-based ONN proposals.

Funding

Sharif University of Technology; Iran National Science Foundation (98012500).

Disclosures

The authors declare no conflicts of interest.

References

1. T. Yan, J. Wu, T. Zhou, H. Xie, F. Xu, J. Fan, L. Fang, X. Lin, and Q. Dai, "Fourier-space Diffractive Deep Neural Network," Phys. Rev. Lett. **123**(2), 023901 (2019).
2. D. Mengu, Y. Luo, Y. Rivenson, and A. Ozcan, "Analysis of Diffractive Optical Neural Networks and Their Integration With Electronic Neural Networks," IEEE J. Sel. Top. Quantum Electron. **26**(1), 1–14 (2020).
3. X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, Y. Luo, M. Jarrahi, and A. Ozcan, "All-optical machine learning using diffractive deep neural networks," Science **361**(6406), 1004–1008 (2018).
4. Y. Shen, N. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and M. Soljacic, "Deep learning with coherent nanophotonic circuits," Nat. Photonics **11**(7), 441–446 (2017).
5. T. W. Hughes, M. Minkov, Y. Shi, and S. Fan, "Training of photonic neural networks through in situ back-propagation and gradient measurement," Optica **5**(7), 864–871 (2018).
6. I. Williamson, T. Hughes, M. Minkov, B. Bartlett, S. Pai, and S. Fan, "Reprogrammable Electro-Optic Nonlinear Activation Functions for Optical Neural Networks," IEEE J. Sel. Top. Quantum Electron. **26**(1), 1–12 (2020).
7. M. Y. S. Fang, S. Manipatruni, C. Wierzynski, A. Khosrowshahi, and M. R. Deweese, "Design of optical neural networks with component imprecisions," Opt. Express **27**(10), 14009–14029 (2019).
8. J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran, and W. H. P. Pernice, "All-optical spiking neurosynaptic networks with self-learning capabilities," Nature **569**(7755), 208–214 (2019).
9. J. Chang, V. Sitzmann, X. Dun, W. Heidrich, and G. Wetzstein, "Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification," Sci. Rep. **8**(1), 12324 (2018).
10. S. Colbburn, Y. Chu, E. Shilzerman, and A. Majuumdar, "Optical frontend for a convolutional neural network," Appl. Opt. **58**(12), 3179–3186 (2019).
11. E. Khoram, A. Chen, D. Liu, L. Ying, Q. Wang, M. Yuan, and Z. Yu, "Nanophotonic media for artificial neural inference," Photonics Res. **7**(8), 823–827 (2019).
12. A. Tait, T. de Lima, E. Zhou, A. Wu, M. Nahmias, B. Shastri, and P. Prucnal, "Neuromorphic photonic networks using silicon photonic weight banks," Sci. Rep. **7**(1), 7430 (2017).
13. R. Hamerly, L. Bernstein, A. Sludds, M. Soljačić, and D. Englund, "Large-Scale Optical Neural Networks Based on Photoelectric Multiplication," Phys. Rev. X **9**(2), 021032 (2019).
14. J. Bueno, S. Maktoobi, L. Froehly, I. Fischer, M. Jacquot, L. Larger, and D. Brunner, "Reinforcement learning in a large-scale photonic recurrent neural network," Optica **5**(6), 756 (2018).
15. Y. Zuo, B. Li, Y. Zhao, Y. Jiang, Y. Chen, P. Chen, G. Jo, J. Liu, and S. Du, "All-optical neural network with nonlinear activation functions," Optica **6**(9), 1132 (2019).
16. Q. Zhang, H. Yu, M. Barbiero, B. Wang, and M. Gu, "Artificial neural networks enabled by nanophotonics," Light: Sci. Appl. **8**(1), 42 (2019).
17. M. Miscuglio, A. Mehrabian, Z. Hu, S. I. Azzam, J. George, A. V. Kildishev, M. Pelton, and V. J. Sorger, "All-optical nonlinear activation function for photonic neural networks," Opt. Mater. Express **8**(12), 3851–3863 (2018).
18. A. Backer, "Computational inverse design for cascaded systems of metasurface optics," Opt. Express **27**(21), 30308 (2019).
19. Z. Wu, M. Zhou, E. Khoram, B. Liu, and Z. Yu, "Neuromorphic metasurface," Photonics Res. **8**(1), 46 (2020).
20. J. Weng, Y. Ding, C. Hu, X. Zhu, B. Liang, J. Yang, and J. Cheng, "Meta-neural-network for Realtime and Passive Deep-learning-based Object Recognition," arXiv:1909.07122 (2019).
21. Z. Wang, T. Li, A. Soman, D. Mao, T. Kananen, and T. Gu, "On-chip wavefront shaping with dielectric metasurface," Nat. Commun. **10**(1), 3547 (2019).

Optics EXPRESS

22. S. AbdollahRamezani, K. Arik, A. Khavasi, and Z. Kavehvash, "Analog computing using graphene-based metalines," Opt. Lett. **40**(22), 5239 (2015).
23. S. Ma, R. Bassily, and M. Belkin, "The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning," arXiv:1712.06559 (2017).
24. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proc. IEEE **86**(11), 2278–2324 (1998).
25. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv:1412.6980 (2014).
26. S. Pai, I. A. D. Williamson, T. W. Hughes, M. Minkov, O. Solgaard, S. Fan, and D. A. B. Miller, "Parallel fault-tolerant programming of an arbitrary feedforward photonic network," arXiv:1909.06179v1 (2019)