High-Speed Multi-Layer Convolutional Neural Network Based on Free-Space Optics

Hoda Sadeghzadeh and Somayyeh Koohi

***Abstract*—Convolutional neural networks (CNNs) are at the heart of several machine learning applications, while they suffer from computational complexity due to their large number of parameters and operations. Recently, all-optical implementation of the CNNs has achieved many attentions, however, the recently proposed optical architectures for CNNs cannot fully utilize the tremendous capabilities of optical processing, due to the required electro-optical conversions in-between successive layers. To implement an all-optical multi-layer CNN, it is essential to optically implement all required operations, namely convolution, summation of channels' output for each convolutional kernel feeding the nonlinear unit, nonlinear activation function, and finally, pooling operations. Considering the lack of multi-layer photonic CNN implementation, in this paper, we explore a fully-optical design for implementing successive convolutional layers in an optical CNN. As a proof of concept, and without loss of generality, we considered two successive optical layers in the proposed network, named as 2L-OPCNN, for comparative studies against electrical counterpart and single optical layer CNN. Our simulation results confirm nearly the same accuracies for classifying images of Kaggle Cats and Dogs challenge, CIFAR-10, and MNIST datasets, compared to the electrical counterpart, as well as improved accuracies compared to single optical layer CNN.**

***Index Terms*—All-optical neural network, deep convolutional neural network, high performance neural network, image classification, optical correlator.**

I. Introduction

INCREASING capability of machine learning technologies and artificial neural network, specifically deep neural networks, has garnered great progress in a variety of applications, including medicine [1], signal processing [2], and many more. In particular, convolutional neural networks (CNNs) enhance performance of computer vision applications, namely image classification [3] and pattern recognition [4].

Indeed, CNNs are at the heart of several machine learning applications [5], but they suffer from computational complexity due to their large number of parameters and operations [6], [7]. As a result, memory usage, power and energy consumption, and computational delay increase in CNNs [6], [7]. Therefore, there is a need of parallelism in software implementation of these networks for processing huge datasets [7]. Although graphical processing units (GPUs) provide parallelism for implementing computational tasks [8], real-time inference is not easily achievable due to their computation time and energy-hungry problems [7]. To date, photonic platform and optical neural networks (ONNs) are an appropriate solution to overcome various drawbacks of electrical implementation [9], [10]. Ultra-broad bandwidth, high interconnection and inherent parallelism capability are the key advantages of optical processing technology [11]–[14] which offers task execution at the speed of light [15]. Recently, some new interests have been activated in developing ONN, and new ONNs have been proposed for implementing photonic multilayer perceptrons (MLPs) [10], [16]–[18] as well as photonic CNNs [6], [7], [19]–[21]. It should be noted that some implementation is based on integrated setup [22], [23], while the others are based on free-space one [19], [20], [24], [25]. Although the integrated structures offer lower power consumption and lower area in comparison with the free-space structures, they suffer from reconfigurability and scalability problems, as well as fabrication challenges. In this manner, free-space optical design is addresses in this paper.

To implement an all-optical multi-layer CNN, it is essential to optically implement all required operations, namely convolution, summation of channels' output for each convolutional kernel feeding the nonlinear unit, nonlinear activation function, and finally, pooling operation. 4f optical correlator is a common architecture in free-space optics, and is utilized by recent studies, such as [5]–[7], [20], [26], as the optical correlator to perform convolution operation in a negligible time [27]. Authors in [6] utilized diffractive optical elements to implement optical correlator for hybrid optical-electronic convolutional neural networks, but they introduced neither optical nonlinear activation function nor optical pooling layer. Moreover, using grayscale input images and one layer of optical convolution, they avoided summation of channels' output for each convolutional kernel. As another usage of 4f system, Colburn *et al.* [7] proposed an optical frontend for AlexNet [3] utilizing metasurface optics to implement array of 4f optical correlator to perform all convolution operations of the first layer. Moreover, they included square nonlinearity at the end of first layer representing photodetector nonlinearity, while other operations were implemented electronically rather than optically.

Authors in [28] proposed a six-successive–layer design of ONN with no nonlinear activation function in-between. Also, they included neither electrical nor optical pooling layer in their

Manuscript received May 24, 2022; accepted June 1, 2022. Date of publication June 8, 2022; date of current version June 17, 2022. *(Corresponding author: Somayyeh Koohi.)*

The authors are with the Sharif University of Technology, Tehran 11365-11155, Iran (e-mail: hoda.sadeghzade@gmail.com; koohi@sharif.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/JPHOT.2022.3180675, provided by the authors.

Digital Object Identifier 10.1109/JPHOT.2022.3180675

ONN architecture. However, they used Sotftmax activation at the end of each 4f system with no optical design proposal to achieve Softmax activation.

Author in [20] proposed an optronic convolutional neural network utilizing a lenslet array of 4f system to perform convolution operations. They proposed an optical summation of channels' output for each convolutional kernel by modulating additional phase shift on each kernel. Moreover, they adopted strided convolution, as the replacement of pooling layer, taking advantages of a 4f optical correlator and a demagnification lens system. Finally, the back-end sCMOS camera is assumed to simulate the electrical nonlinear activation function. It should be noted that the demagnified pooling suffers from two drawbacks: a) the pooling operation actually happens in camera, where the output of the optical system gets demagnified before getting into the camera. Therefore, such a design introduces challenges for implementing two or more successive optical layers and there is always a need of electro-optical conversion between successive layers (due to the usage of camera in-between), and b) even assuming design changes to implement successive all-optical layers, using the demagnifier element as a pooling layer reduces the image sizes exponentially at each layer. And hence, it becomes more and more difficult to perform operations using spatial light modulators (SLMs) or capture the final image using CCD.

In [19], array of 4f optical correlators as the optical convolution layer, saturable absorption as the optical nonlinearity unit, and convolution with pinhole masks as the optical pooling layer have been designed for the first convolutional layer of AlexNet. However, to facilitate optical implementation of CNNs in more than one layer, optical summation of channels' output for each convolutional kernel should be provided to feed the optical nonlinearity units. Moreover, although blurring the transmitted image by passing through a low pass filter (i.e., a pinhole mask) simulates the average pooling operation, an efficient AlexNet architecture utilizes a max pooling layer. Therefore, either optical implementation of max pooling operation or an optical function with similar behavior as a max pooling unit should be provided to optimize the network classification accuracy.

In sum, considering all the aforementioned ingredients, there is a lack of optical multi-layer CNNs capable of optically implementing all successive operations, namely convolution, summation of channels' output for each convolutional kernel feeding the nonlinear unit, nonlinear activation function, and finally, pooling layer. It is worth noting that multi-layer photonic CNNs, with no electro-optical conversion between successive layers, can reduce power consumption of electrical CNNs. Moreover, as discussed in [7], in conventional CNNs (such as AlexNet [3]), the most time-consuming layers are the first and the second ones. Therefore, optical implementation of all required operations facilitates optical implementation of successive layers. In this manner, an optical multi-layer CNN allows considerable speed up over traditional counterparts

In this paper, we introduce a fully-optical design for implementing successive convolutional layers in an optical CNN. Without loss of generality, as a case study, we simulated two optical layers, as a proof of concept for concatenating optical layers. The proposed architecture, named as 2L-OPCNN (i.e., Two Layer Optical CNN), would be implemented in free-space optics, taking advantages of 1) array of 4f optical correlators (to implement optical convolution), 2) phase shifting of kernels (to implement summation of channels' output), 3) saturable absorption nonlinearities (as optical nonlinearities), and 4) array of 4f optical correlators (to implement optical depth-wise convolution as a replacement of max pooling). It is worth noting that replacement of max pooling by a convolution layer was previously introduced in [29] for electrical CNNs. However, adoption of 4f optical correlator to achieve as high accuracy as the max pooling layer is investigated in this paper. In other words, the significant novelties of this work are designing a generalized optical pooling solution, to be implemented by a trainable optical convolution layer, rather than the max pooling layer, and concatenation of all operational optical blocks to implement an all-optical CNN with no electro-optical conversion between successive layers. In addition, thanks to the great properties of optical implementation, the proposed solution provides speedup, as well as negligible power consumption in implementing any optical CNN which are inevitable cornerstone in electrical CNN. It is worth noted that optical implementation of CNN can be used in real-life applications, such as processing of large biological data sequences. As explained in [30], while designing sequence alignment tool for biological sequences (i.e., DNAs, RNAs, or proteins) is a debatable issue, adopting optical correlators can speed up the computation time by 81% against the electrical counterparts.

Based on the information elaborated upon, in this paper, we design a fully-optical multi-layer CNN. The rest of the paper is organized as follows. Section II presents the 2L-OPCNN structure in details, and explores the optical implementation of each operation. Section III presents the simulation environment and the accuracy evaluation of the proposed photonic CNN. In Section IV, we discuss the speedup achieved by the 2L-OPCNN architecture against its electrical counterpart. Section V addresses the scalability analysis in term of the area consumption and alignment noises. Section VI represents the power analysis of 2L-OPCNN, and finally, Section VII concludes the paper and presents the future work.

II. 2L-OPCNN

A generic CNN comprises of convolutional layers, nonlinear activation layers, pooling layers, fully connected layers and an output layer [31]. Without loss of generality, this paper utilizes AlexNet [3] as a famous CNN architecture with several successive convolutional layers and FC layers. Specifically, it consists of five convolutional layers and three fully connected layers. The first and the second layers provide successive operations of convolution, Rectified linear unit (ReLU) activation function, local response normalization (LRN), and max pooling operation. The third and the fourth layers include convolution and ReLU, and finally the fifth layer comprises of convolution, ReLU, and max pooling layers. It should be noted that the initial layers of the CNN consume the most total run time compared to the successive layers [7]. Specifically, the first and the second layer of the AlexNet network consume more than half of the total run time, while the second layer has the highest computational cost by consuming 37.6% of the total run time [7]. In this manner, optical implementation of the time-consuming initial layers has motivated various researches to propose optical networks.

In this section, the details of the proposed optical network, namely 2L-OPCNN, are provided by presenting optical implementation of the first and the second layers of AlexNet [3]. It is worth mentioning that although 2L-OPCNN takes advantages of two optical layers, the proposed optical convolutional layer can be utilized in any other convolutional neural network. Fig. 1 shows the building blocks of 2L-OPCNN architecture. It is worth noting that although the first and the second layers of 2L-OPCNN are implemented optically, providing optical implementation of two successive layers proves feasibility of all-optical multi-layer CNN architecture. Moreover, due to various design considerations in free-space optics, following assumptions are made for the first and the second layers, as shown in Fig. 1:

I) Convolution operations are implemented by utilizing 4f optical correlators with Fourier lenses and stride value of 1, rather than 4, due to the continuous Fourier transform properties,

II) Bias terms are omitted to simplify the optical implementation,

III) Saturable absorber (SA) optical nonlinearity [32] is adopted instead of ReLU,

IV) As a result of input normalization for increasing the performance of SA functionality, there is no requirement of LRNs, and so, the LRNs do not affect the classification accuracy [19].Therefore, we omitted them in both layers, and finally,

V) As a key advantage of 2L-OPCNN, we replace the max pooling operation by a depthwise convolution layer. Moreover, since there is no way to apply overlapping in free-space optics, we optically implement pooling units by 4f optical correlators with specific filter size for each dataset, stride value of 1 (considering continuous Fourier transform), and a downsampling layer, following the photodector's square nonlinearity response, at the end of the second layer.

A. Optical Implementation of Fourier Convolution

Cross-correlation measures the similarity between two signals/sequences [19]. On the other hand, convolution measures the effect of one signal on the other one. It is worth noting that the mathematical calculations of correlation and convolution are similar in the time domain, except that the signal is not reversed, before the multiplication process, in the case of correlation. In other words, for symmetric filters, the outputs of two operations would be similar. It is worth noting that while performing convolution operation, convolution of image and kernel, and convolution of image and the reversed kernel lead to the similar outputs. Concluding aforementioned discussion, we can state that convolution and correlation operations achieve similar outputs for CNN architectures. Furthermore, we can calculate the cross-correlation operation in the frequency domain, as in (1), where F is the Fourier transform, G(u,v) is the Fourier transform of g(u,v), S*(u,v) is the complex conjugate function of s(u,v), and finally, c(x,y) shows the 2D correlation of the two functions [27].

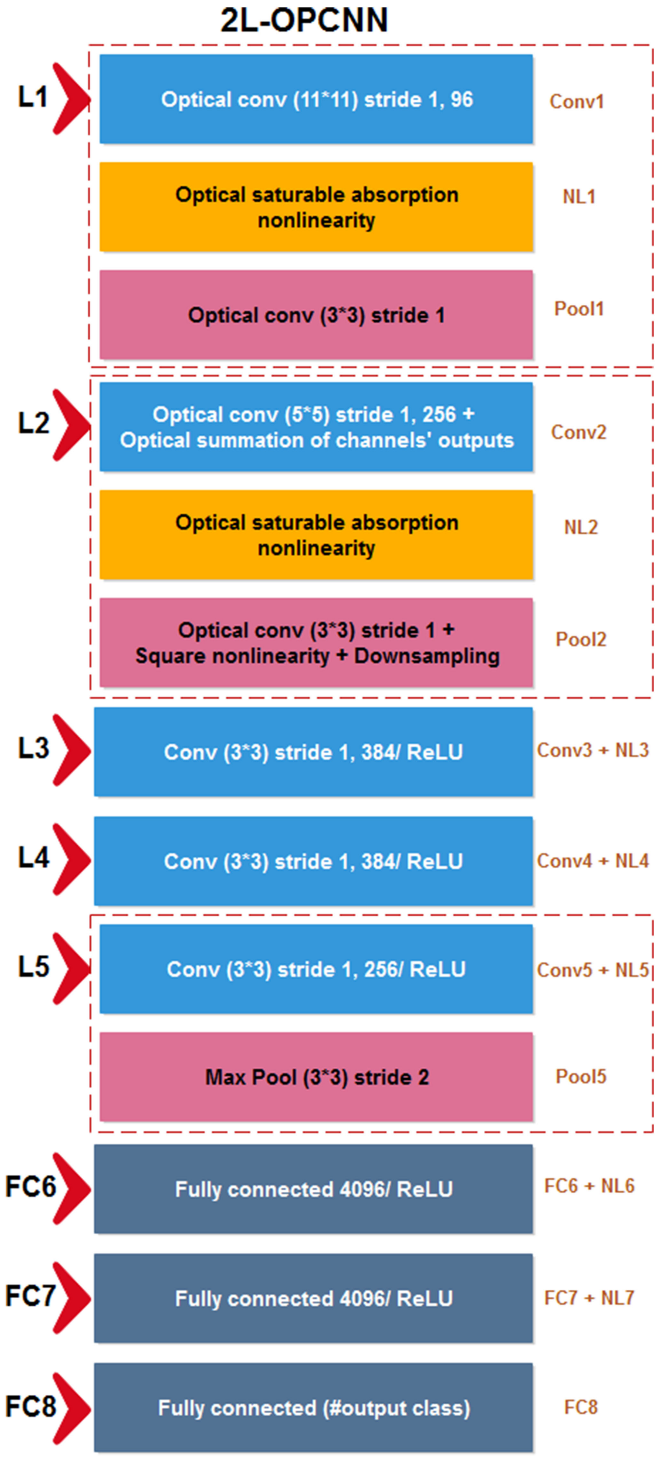
2L-OPCNN
L1
Optical conv (11*11) stride 1, 96
Conv1
Optical saturable absorption nonlinearity
NL1
Optical conv (3*3) stride 1
Pool1
L2
Optical conv (5*5) stride 1, 256 + Optical summation of channels' outputs
Conv2
Optical saturable absorption nonlinearity
NL2
Optical conv (3*3) stride 1 + Square nonlinearity + Downsampling
Pool2
L3
Conv (3*3) stride 1, 384/ ReLU
Conv3 + NL3
L4
Conv (3*3) stride 1, 384/ ReLU
Conv4 + NL4
L5
Conv (3*3) stride 1, 256/ ReLU
Conv5 + NL5
Max Pool (3*3) stride 2
Pool5
FC6
Fully connected 4096/ ReLU
FC6 + NL6
FC7
Fully connected 4096/ ReLU
FC7 + NL7
FC8
Fully connected (#output class)
FC8

Fig. 1. Schematic of 2L-OPCNN architecture. For each layer, the corresponding operations, filter size, stride value, and the number of kernels is specified.

$$c\left(x,y\right)=F\left\{G\left(u,v\right)S^{*}\left(u,v\right)\right\} \qquad (1)$$

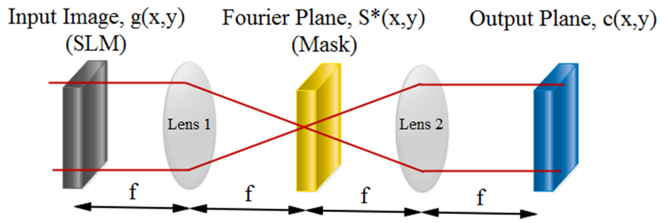

Fig. 2. Schematic of a typical 4f optical correlator.

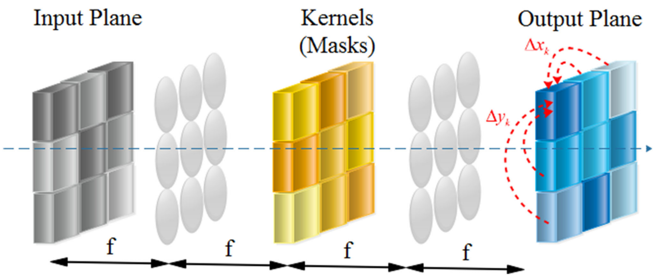

Fig. 3. An array-based convolutional layer with adding phase shifts on kernels.

The above functionality can be realized by a 4f system in free-space optics. 4f system is a common architecture to implement optical correlation in free-space optics which is based on the Vanderlugt setup [33] and comprises of two Fourier lenses each with focal length of f. Fig. 2 shows the schematic of a typical 4f system.

As shown in Fig. 2 an input image is formed by modulating amplitude, phase, or polarization of light beams by a SLM. Afterwards, light transmission through the first lens (Lens 1) results in Fourier transform of the input image. While, the second image is Fourier transformed in the offline manner, and the corresponding mask is located on the Fourier plane. Multiplication of two transformed images is occurred in the Fourier plane as well. By light propagation through the second lens (Lens 2), the proposed setup performs Fourier transform once again to achieve the correlation of two images. The resultant correlation output either propagates to the next optical layers for more optical processing (as discussed in this section) or is captured by a camera [6], [7] and converted to electrical signals for further electrical processing.

B. Optical Summation of Channels' Outputs for Each Convolutional Kernel

For parallel implementation of all convolution operations, we use array of 4f optical correlators for the first and the second convolution layers of 2L-OPCNN. It is worth mentioning that each kernel extracts its related feature map by performing the convolution operation [20]. Moreover, for optical CNNs with more than one convolutional layer, the RGB input images require three channels for each kernel of the first convolutional layer, and so, optical summation of channels' output for each convolutional kernel is required before feeding data to the optical nonlinearity units in every convolutional layer. However, in the case of grayscale input images, each kernel of the first convolutional layer has one channel, and so, summation of channels' output for each convolutional kernel is not required in the first convolutional layer, while it is required for the 2nd to the last convolutional layers. We would like to emphasize that assuming the grayscale input images in this paper (as details are explored in Section III.A), only one channel exists for each kernel in the first layer, and so, no summation operation is required in the first layer. However, the second layer consists of 96 channels for each kernel, and hence, optical summation of channels' output for each convolutional kernel should be performed before feeding data to the optical nonlinearity units.

Optical implementation of summation can be realized by the means of additional modulating of phase shift on each kernel based on shift theorem in Fourier Optics [20]. By tiling the input images $I_{in}(x_k, y_k)$ and kernels $Kernel(x_k/\lambda f, y_k/\lambda f)$ on SLMs, where λ is the wavelength of the free-space light, f is the focal length of the lens, and x_k and y_k represent the pixel coordinates of the kth input image, and taking advantages of an array of 4f systems, convolution of each kernel with its corresponding input can be realized separately. To sum the 96 convolutions' outputs corresponding to 96 channels of each kernel, in the second optical layer, we modulate each kernel value by an additional phase shift of $\phi(\Delta x_k, \Delta y_k)$, where, $\Delta x_k$and $\Delta y_k$are defined according to the required pixels shift of the corresponding kernel image to locate it in its target position. The required mathematical computations are presented as follows [20]:

$$Kernel_{in}\left(\frac{x_k}{\lambda f}, \frac{y_k}{\lambda f}\right) = Kernel\left(\frac{x_k}{\lambda f}, \frac{y_k}{\lambda f}\right).e^{j\varphi(\Delta x_k, \Delta y_k)} \tag{2}$$

where, x_k and y_k represent the pixel coordinates of the kth input image fed to the summation unit, (.) represents 2D element-wise product, and finally, φ is a phase gradient profile applied to the kth kernel. By substituting the modulated kernel of (2) in following (3) (which represents the mathematical operation of the 4f system), and going through mathematical operations, the result can be formulated as (7) [20].

$$I_{out}(x_k, y_k) = F^{-1}\left\{F[I_{in}(x_k, y_k)].Kernel\left(\frac{x_k}{\lambda f}, \frac{y_k}{\lambda f}\right)\right\}$$

$$k = 1, 2, \ldots, N \tag{3}$$

$$I'_{out}(x_k, y_k) = F^{-1}\left\{F[I_{in}(x_k, y_k)].Kernel_{in}\left(\frac{x_k}{\lambda f}, \frac{y_k}{\lambda f}\right)\right\} \tag{4}$$

$$= F^{-1}\left\{F[I_{in}(x_k, y_k)].Kernel\left(\frac{x_k}{\lambda f}, \frac{y_k}{\lambda f}\right). e^{j\varphi(\Delta x_k, \Delta y_k)}\right\} \tag{5}$$

$$= F^{-1}\{F[I_{out}(x_k, y_k)].e^{j\varphi(\Delta x_k, \Delta y_k)}\} \tag{6}$$

$$= I_{out}(x_k - \Delta x_k, y_k - \Delta y_k)k = 1, 2, \ldots, N \tag{7}$$

Based on the shift theorem, adding these phase shifts on each kernel causes the whole output feature maps to be superimposed in a specific position, as shown in Fig. 3. In general, when a phase gradient is applied to the phase profile of the SLM, the reflected light field is deflected accordingly, and hence, each output field $I_{out}(x_k,y_k)$ deflects to the target position as well. Therefore, summation of channels' output for each convolutional kernel can be performed optically, according to (8).

$$I'_{out}(x,y) = \sum_{k=1}^{N} I'_{out}(x_k,y_k) \tag{8}$$

To determine the new phase profile of each kernel, we can combine (6) and (7) as (9).

$$\begin{aligned} &F^{-1}\{F[I_{out}(x_k,y_k)].e^{j\varphi(\Delta x_k,\Delta y_k)}\} \\ &= I_{out}(x_k-\Delta x_k, y_k-\Delta y_k)\; k=1,2,\ldots,N \end{aligned} \tag{9}$$

Or by performing a Fourier transform we can rewrite it as follows:

$$F\,[I_{out}(x_k,y_k)].e^{j\varphi(\Delta x_k,\Delta y_k)} = F[I_{out}(x_k-\Delta x_k, y_k-\Delta y_k)] \tag{10}$$

$$e^{j\varphi(\Delta x_k,\Delta y_k)} = \frac{F[I_{out}(x_k-\Delta x_k, y_k-\Delta y_k)]}{F[I_{out}(x_k,y_k)]} \tag{11}$$

Measuring I_{out} for the k^{th} kernel, by assuming pre-specified Δx_k and Δy_k according to the desired pixels shift of each image, we can extract exact value of phase shift φ for each kernel.

C. Optical Saturable Absorption Nonlinearity

One of the most challenging part of implementing the ONNs is realizing physical optical nonlinearity [25]. Saturable absorption nonlinearity is an all-optical nonlinearity which can be performed in free-space optics by passing light through an atomic vapor. An atomic vapor cell is a glass cell containing a specific gas, which represents specific absorption spectrum, and it provides nonlinear relation between its input and output propagated lights [25]. (12) and (13) represent the mathematical model of saturable absorber (SA) [32] and its derivative, presuming a real-valued E-field (otherwise the square term should be the square of the absolute value, i.e., $|E|^2$):

$$E_{P,out} = g\left(E_{P,in}\right) = \exp\left(-\frac{\alpha_0/2}{1+E_{P,in}^2}\right) E_{P,in} \tag{12}$$

$$g'\left(E_{P,in}\right) = \left[1 + \frac{\alpha_0 E_{P,in}^2}{\left(1+E_{P,in}^2\right)^2}\right] \exp\left(-\frac{\alpha_0/2}{1+E_{P,in}^2}\right) \tag{13}$$

where, $E_{P,in}$ is the input signal of SA, α_0 is the resonant optical depth, $g(E_{P,in})$ represents the nonlinear output, and $g'(E_{P,in})$ is the derivative output. Fig. 4(a) and 4(b) represent the input-output transmission function of SA and its derivate assuming optical depth of 20, and 30, respectively. It is worth mentioning that SA has two processing regions: (i) nonlinear and (ii) linear regions, and so, based on the input intensity, it can behave diversely. Also, by increasing the resonant optical depth of SA, it operates in wider nonlinear region. It should be noted that to evaluate the temporal behavior of SA, the atomic model of SA can be formulated by a system of coupled differential equations, so-called as rate equations [34], the details of which are discussed in [19].

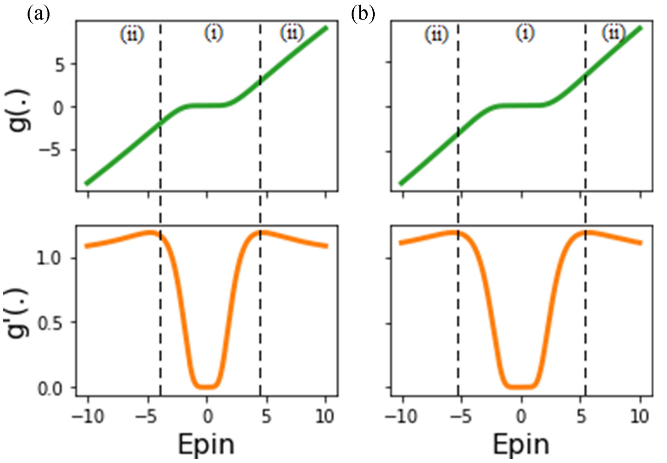

Fig. 4. Input-output functionality of SA (green color) and its derivative (orange color) assuming resonant optical depth of (a) 20 and (b) 30.

D. Optical Max Pooling

Pooling layers, contributing in most CNNs, progressively reduce the spatial size of the input representation, as well as reduce the numbers of parameters in each layer to speed up the network computation. Providing spatial invariance property, pooling layers reduce overfitting in the neural networks, among which, the average pooling and the max pooling layers are the most popular ones [35]. However, although a few recent studies propose an optical demagnification system as the pooling layer [20], or adoption of pinhole as an optical average pooling layer [19], there is a lack of proper implementation of the max pooling layer within a free-space optical setup [32].

Recently, a few studies [29], [35] proposed utilization of the convolution layers instead of the max pooling layers to enhance the accuracy of electrical CNN. These study prove that the max pooling operation is mathematically equivalent to the convolution operation followed by an appropriate nonlinear activation function, such as ReLU [29], [35]. While all these architectures are proposed for the electrical CNN, in this paper, we present an optical solution for this approach. As shown in Fig. 5 , we can deduce that a trainable convolutional layer with the appropriate kernels can resemble the max pooling operation. Therefore, we propose to substitute the max pooling layer with a convolutional layer with stride value more than 1. Based on the above explanation, the max pooling operation is mathematically equivalent to the convolution operation followed by a nonlinear activation function. However, to reduce the complexity of the optical setup, we proposed to remove the nonlinear activation function while implementing the convolution operation to resemble the max pooling operation. Therefore, due to the omission of the nonlinear activation function, we cannot claim that the implemented convolution is mathematically equivalent to the

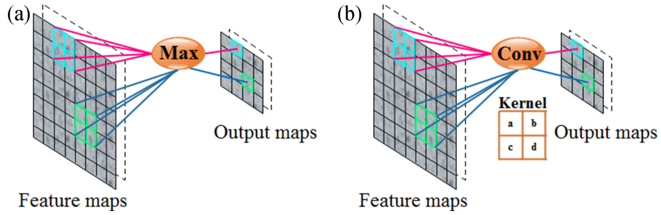

Fig. 5. (a) Max pooling operation and (b) A trainable convolutional layer with appropriate kernels can achieve the same classification accuracy as the max pooling operation. Parameters a, b, c, and d are kernel's pixels value which are determined though training procedure.

max pooling. However, as a key advantage of the trainable convolution operation and the learning process, we could achieve similar accuracies by the convolution operation as the pooling layer, compared to the max pooling layer. For more clarity, as an example, a max pooling layer with filter size of 2 and stride value of 2 can be substituted with a convolutional layer with the same filter size and stride value. In this case, the numbers of output and input channels are not altered, and hence, a depthwise convolution operation can be adopted.

Based on above discussion, in this paper, convolution-based pooling layers as the substitution of max pooling layers are adopted in the optical domain. Specifically, we take advantage of optical convolution operations instead of max pooling operation in the first and the second layers of 2L-OPCNN structure. As discussed in Section II.A, we perform optical convolution using common 4f optical correlators, and since there is no way of implementing stride values larger than one in the optical convolution, we consider downsampling operation within the CCD camera, which captures output of the second layer.

E. Optical Layers Structure of 2L-OPCNN

Figs. 6 and 7 show all-optical design of the first and the second convolutional layers of 2L-OPCNN with arrays of 4f optical correlators for both convolution and pooling layers, as well as arrays of SAs performing nonlinear blocks. As shown in these figures, although similar, the first and the second optical layers differ in two aspects: I) assuming grayscale input images for 2L-OPCNN, all kernels in the first convolution operations within the first layer requires one channel. Therefore, no optical summation is required at the end of the first layer, and II) utilizing CCD camera, performing opto-electrical conversion, at the end of the second layer is inevitable to facilitate further electrical processing.

It is obvious that all training kernels in both layers are determined through an electrical training procedure. Regarding the proposed optical design, for an optical implementation of the test procedure, each positive or negative kernel's value is implemented utilizing a checkerboard pattern of subpixels [7], whose details are discussed in [19] and [36]. Also, we utilize flat lenses based on the metasurfaces, whose phase shift φ_L at each point of (x,y) on the flat lens is calculated as follows [37]:

$$\varphi_L = \frac{2\pi}{\lambda}\left(\sqrt{x^2+y^2+f^2}-f\right) \tag{14}$$

TABLE I
Number of Images for Network Training, Validation and Test

Dataset	#Total images	#Training images	#Validation images	#Test images
Kaggle Cats and Dogs	37.5 K	30 K	2.5 K	5 K
CIFAR-10	60 K	45 K	5 K	10 K
MNIST	70 K	55 K	5 K	10 K

where, λ is the wavelength in free-space, and f is the focal length of the lens. In this case study, for implementing 2L-OPCNN, we assumed λ of 532 nm, f of 3 mm, and lens diameter of 0.57 mm. In this manner, as the 2L-OPCNN architecture utilizes 96 kernels in the first layer, the lens array performing the optical convolution operations within the first convolutional layer constitutes an area of 0.31 cm^2. Moreover, since each optical convolution operation is followed by an SA component, 96 SA components are required to accomplish the nonlinearity operation. At the last stage of the first optical layer, we adopt 96 4f optical correlators aligned with SAs' outputs to implement the required optical pooling operations in a parallel manner. Therefore, the optical pooling layer constitutes an area of 0.31 cm^2 in the first convolutional layer. In the second convolutional layer, we apply 256 kernels in an array-based structure, and so, area of 79.85 cm^2 is required for the lens arrays implementing both the convolution and the pooling operations in a parallel manner, while 256 SA components are utilized in between. Finally, it should be noted that since the cells diameter of SA array is 19 mm or 25.4 mm, as accessible in [38], it consumes less area compared to the lenslet array, and would not limit the scalability of the optical setup.

III. Simulation and Results

In this section, to evaluate the accuracy of 2L-OPCNN, we present different simulation scenarios involving various input datasets. Moreover, the details of training procedure and the resultant classification accuracy for each dataset are described in the following subsections.

A. Datasets

Three classification datasets, namely Kaggle Cats and Dogs [39], CIFAR-10 [40], and MNIST [41] were chosen for performance analysis of 2L-OPCNN. The Kaggle Cats and Dogs includes two classes of RGB images, while the CIFAR-10 includes 10 classes of 32 × 32 RGB images, and finally, the famous MNIST consists of 10 classes of 28 × 28 grayscale images. Table I shows total number of images in each dataset and the corresponding numbers of training, validation, and test images considered in our simulations. It is worth mentioning that based on the 2L-OPCNN input architecture, all input images were resized to 227 × 227 pixels, and the RGB input images were converted to the grayscale ones.

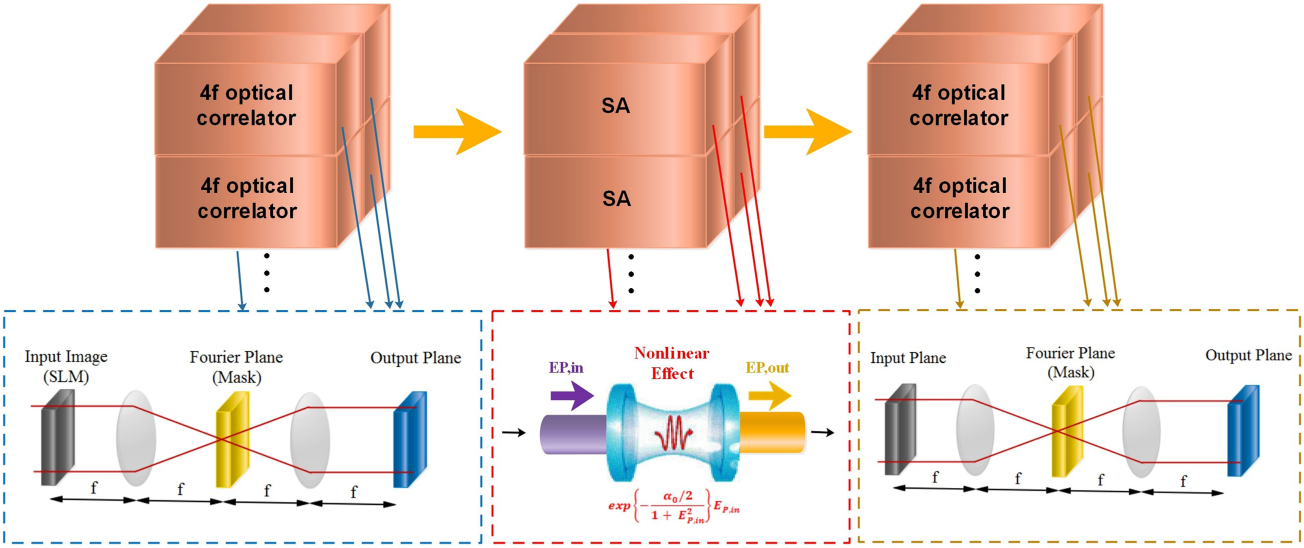

Fig. 6. All-optical implementation of the first convolutional layer with array-based structure.

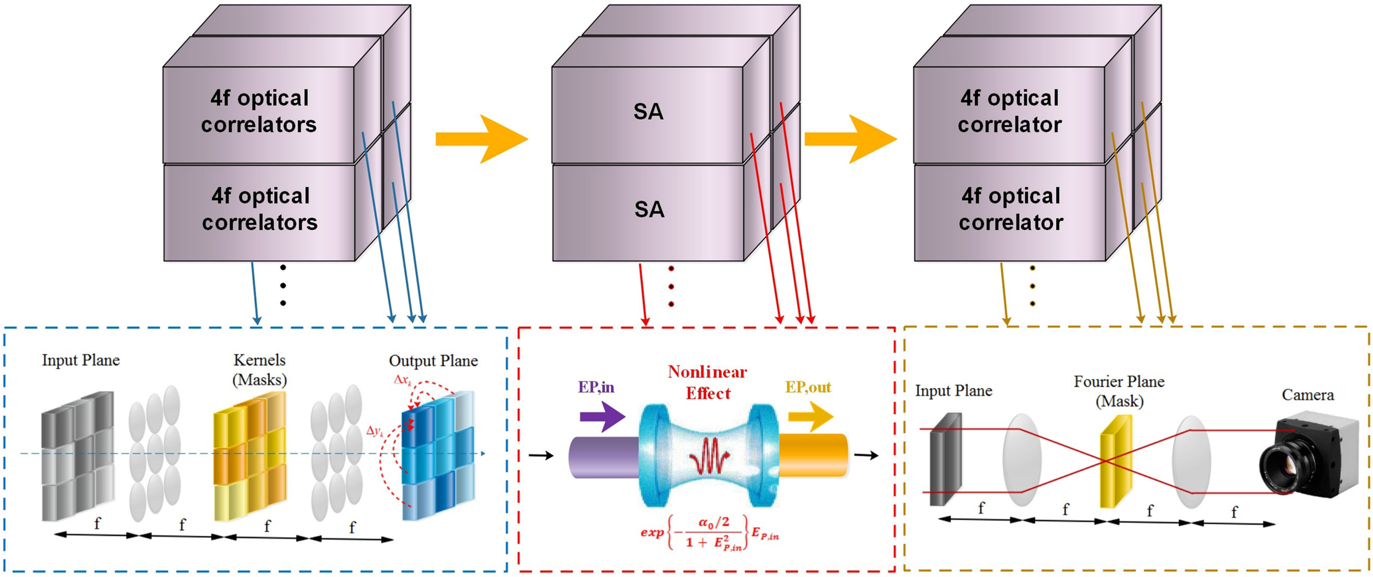

Fig. 7. All-optical implementation of the second convolutional layer with array-based structure.

B. Comparative Simulation Studies

In this study, three simulation scenarios are considered, namely as AlexNet, 2L-OPCNN (Ground Truth), and 2L-OPCNN (wave-optics); where, AlexNet represents the conventional AlexNet structure [3], 2L-OPCNN (Ground Truth) is simulated to measure the accuracy of 2L-OPCNN considering behavioral simulation and analytical models of optical components, and finally, 2L-OPCNN (wave-optics) represents the corresponding wave-optics simulation by extending the wave optics-based code evolved in [7].

In 2L-OPCNN (Ground Truth), we applied the following main modifications on AlexNet [3]: I) the corresponding convolution operation is performed in the optical domain by a Fourier domain multiplication. Considering the continuous Fourier transform, we utilized a stride of 1, instead of 4 in AlexNet [3], II) the SA nonlinearity, instead of ReLU, is applied for the first and the second layers of 2L-OPCNN, and finally III) we replaced the max pooling operation of both the first and the second layers by the trainable convolution operations. 2L-OPCNN (wave-optics) considers wave optics simulations of optical component in the Fourier domain. For this purpose, we included the fast Fourier transform (FFT) algorithm, angular spectrum propagator, and complex-valued masks in Tensorflow-Python framework. For a fair comparison, all scenarios were evaluated by grayscale input images.

TABLE II
STRUCTURAL DETAILS OF FIRST AND SECOND LAYERS STRUCTURE OF ALEXNET AND 2L-OPCNN

Name	Convolution1 + Nonlinearity1	Pooling1	Convolution2 + Nonlinearity2	Pooling2	Nonlinearity
AlexNet	Conv2D(11*11) and stride of 4 + bias+ ReLU +LRN	Max-pool(3*3) and stride of 2	Conv2D(5*5) and stride of 1 + bias+ ReLU +LRN	Max-pool(3*3) and stride of 2	-
2L-OPCNN (Ground Truth)	Conv2D(11*11) and stride of 1 + SA	Conv2D(d*d) and stride of d (d = 3 or 4)	Conv2D(5*5) and stride of 1 + SA	Conv2D(3*3) and stride of 3	Sqnl
2L-OPCNN (wave-optics)	Wave-optics Conv + SA	Wave-optics Conv	Wave-optics Conv + SA	Wave-optics Conv	Sqnl + Downsampling

TABLE III
SIMULATION RESULTS COMPARISON OF 2L-OPCNN AND ALEXNET

Dataset	AlexNet	2L-OPCNN (GROUND TRUTH)	2L-OPCNN (WAVE-OPTICS)
Kaggle Cats and Dogs	88.52	88.16	87.40
CIFAR-10	78.50	77.41	74.15
MNIST	99.12	99.36	99.34

TABLE IV
SIMULATION RESULTS COMPARISON OF 2L-OPCNN AND OP-ALEXNET [19]

Dataset	2L-OPCNN (wave-optics)	OP-AlexNet (OPL1-Conv-SA-AvgPool)[19]
Kaggle Cats and Dogs	87.40	83.76
CIFAR-10	74.15	72.82
MNIST	99.34	99.25

To choose the best filter sizes for the convolution-based max pooling layers, numerous simulation scenarios were performed, as the details are provided in Section I of supplementary material. In this regard, in the case of Kaggle Cats and Dogs and MNIST datasets, filter size of (3 × 3) and stride of 3 were selected for convolution-based pooling operations of both first and second layers, while for CIFAR-10 dataset filter size of (4 × 4) and stride of 4, and filter size of (3 × 3) and stride of 3 are chosen for the first and the second pooling layers, respectively. Moreover, as discussed in Section II, we drop bias terms and LRN unit of all convolution operations within the first and the second layers of 2L-OPCNN. Finally, it is worth noting that square nonlinearity (Sqnl) representing the nonlinear response of the photodetecror is adopted at final stage of the second layer of 2L-OPCNN structure. Summarizing the above discussion, Table II represents the details of various simulation scenarios.

To achieve the best classification accuracy, we considered weights initialization with a Gaussian distribution assuming std value of $1/\sqrt{fan_{in}}$, where $\mathrm{fan_{in}}$ is computed as $\mathrm{m^2c}$ for the convolutional layer with kernel size of m × m and c input channels. Also, we set the resonant optical depth (α_0) of 20 for SA nonlinearity, and finally, we executed all the aforementioned simulation scenarios on GPU (NVIDIA GeForce GTX). Moreover, to obtain the best parameter values, several training procedures were performed on validation datasets, and finally, the learning rate of 0.001 and batch size of 4 were considered for 2L-OPCNN (wave-optics). The learning rate and batch size of 2L-OPCNN (ground-truth) are provided in the supplementary material. Also, it should be considered that the training procedure is stopped once the rate of change of training accuracies and cross entropies reduces to $<10^{-3}$.

Table III lists the resultant classification accuracies. It is worth mentioning that the final goal of almost all researches [6], [7], [20] proposing OPCNN is to reach a classification accuracy close to that of the electrical counterpart. In this manner, we compared the 2L-OPCNN with its electrical counterpart in terms of classification accuracy. As shown in Table III, for Kaggle Cats and Dogs and MNIST datasets, classification accuracy of the 2L-OPCNN (wave-optics) is nearly similar to that of the AlexNet, while its accuracy is slightly reduced for the CIFAR-10 dataset. In this manner, we can conclude that optical convolution-based pooling layer, by achieving nearly the same classification accuracy as the max pooling layer, can facilitate all-optical implementation of the convolution layers. Fig. 8 represents all 96 output images of the first convolutional layer, and also, all 256 output images of the second convolutional layer by feeding a sample input image from Kaggle Cats and Dogs dataset.

C. Two Layers vs. One Layer Optical Convolution

To evaluate the accuracy of optical summation in-between subsequent optical layers, in this section, the classification accuracy of 2L-OPCNN is compared against that of OP-AlexNet [19] which consists of one optical convolutional layer and subsequent electrical ones.

As shown in Table IV, the number of optical layers surprisingly impacts the classification accuracy. Specifically, while OP-AlexNet's accuracy is 4.76% and 5.68% less than that of AlexNet (shown in Table III) for Kaggle Cats and Dogs and CIFAR-10 datasets, respectively, accuracy of 2L-OPCNN (wave-optics) is improved by 3.64% and 1.33%, respectively, compared to that of OP-AlexNet for Kaggle Cats and Dogs and CIFAR-10 datasets.

As discussed in [19], including a square nonlinearity operation, representing the photodetector responsivity, at the back end of the first optical convolutional layer considerably reduces classification accuracy of OP-AlexNet. However, as reported in Table III, 2L-OPCNN can achieve nearly the same accuracy as AlexNet. Actually, two main reasons should be noted for the classification accuracy enhancement of 2L-OPCNN over OP-AlexNet: I) utilizing optical trainable convolution layer instead of max pooling, which also involves the pooling layer

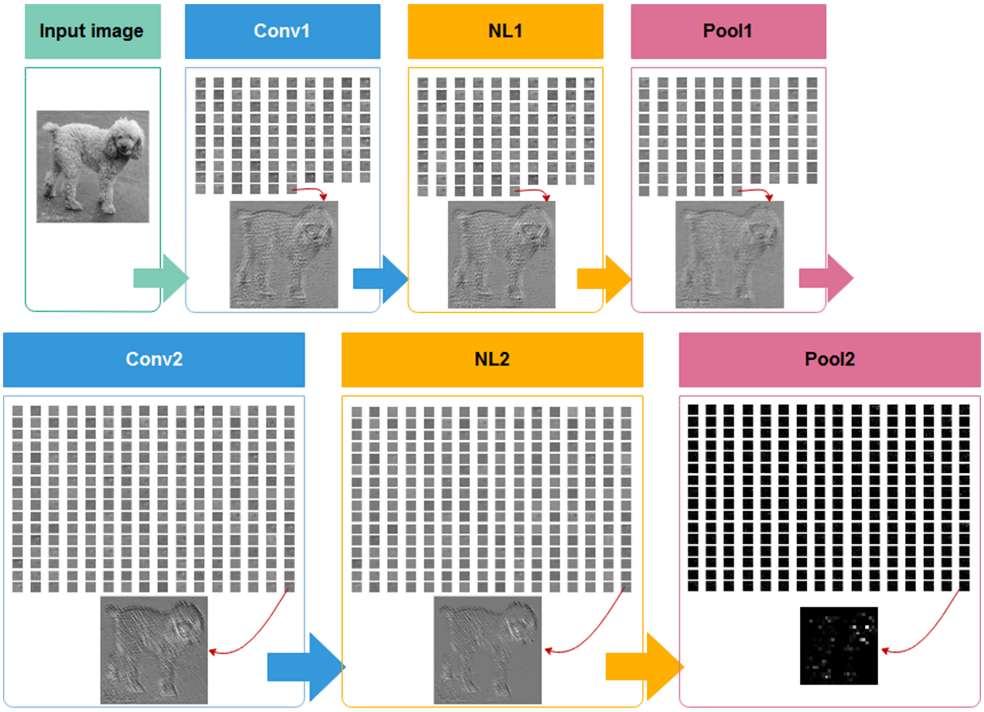

Fig. 8. The output images of each block of first and second layer of 2L-OPCNN.

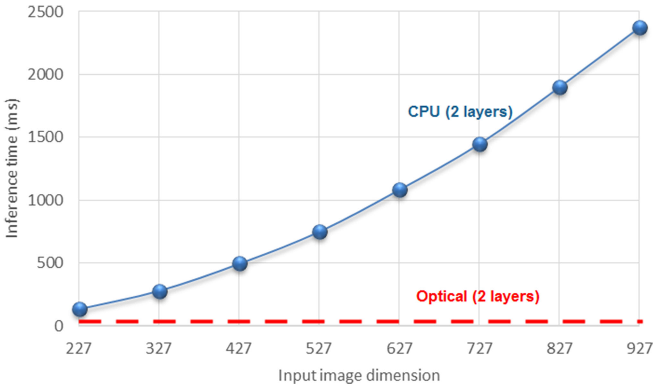

Fig. 9. Comparing the inference time of the optical and electrical implementations of the first two layers of AlexNet in terms of input mage size.

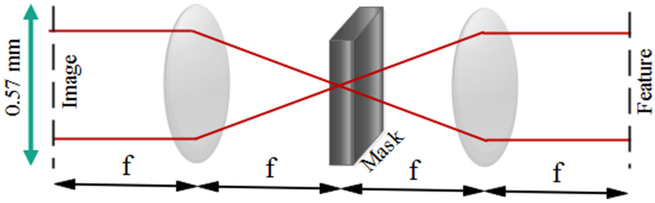

Fig. 10. The side view of a 4f system.

through the training process, compared to the non-trainable max pooling layer, and II) implementing two successive optical layers by applying optical summation. Concatenation of optical layers weakens the negative impact of square nonlinearity on the classification accuracy. In all, utilizing optical summation along with all other optical blocks results in an all-optical design which is a general optical implementation and can be used in any CNN architecture.

IV. Speed Comparison

One of the most fascinating capabilities of optical computing is its high speedup, against the electrical implementation. For a detailed speed comparison, the latency of 2L-OPCNN (wave-optics) can be estimated as follow:

$$2L - Latency = T_{source} + \underbrace{T_{4f\_conv} + T_{SA} + T_{4f\_pool}}_{T_{operation1}} + \underbrace{T_{4f\_conv} + T_{SA} + T_{4f\_pool}}_{T_{operation2}} + T_{camera} + T_{transfer\_data} \tag{15}$$

where, 2L-Latency represents latency of the two optical layers. $\mathrm{T}_{\mathrm{source}}$ represents modulation delay of the input images; and so, considering SLMs with 1 kHz switching frequency [7], $\mathrm{T}_{\mathrm{source}}$ equals 1ms. $\mathrm{T}_{\mathrm{4f^{\cdot}conv}}$ and $\mathrm{T}_{\mathrm{4f^{\cdot}pool}}$ represent the optical propagation delays through the convolution and pooling layers. Considering 4f optical correlators for the aforementioned two layers, $\mathrm{T}_{\mathrm{4f^{\cdot}conv}}$ and $\mathrm{T}_{\mathrm{4f^{\cdot}pool}}$ approximately equal 10 ps [7], [19], which is almost negligible. T_{SA}, as the delay of SA nonlinearity unit, equals 25 ps [19], which is also negligible. $\mathrm{T}_{\mathrm{camera}}$ represents the latency of photodetectors to capture and convert output images to the electrical data. Utilizing high-speed commercial cameras [42], at the speed of 2500 frames per second, the latency of the camera can be estimated as 0.4 ms. Finally, $\mathrm{T}_{\mathrm{transfer^{\cdot}data}}$ is the delay of the communication interface to transmit camera's output to a computer. By considering USB 3.1 Gen2 with a frame rate of 10 Gbit/s and assuming 100 kB image, $\mathrm{T}_{\mathrm{transfer^{\cdot}data}}$ equals 0.08 ms.

TABLE V
COMPARING FORWARD COMPUTATION TIME (MS) OF CONVOLUTIONAL LAYERS OF ALEXNET, 2L-OPCNN AND FIVE OPTICAL LAYERS FOR AN INPUT IMAGE WITH SIZE OF 227×227

Network	L1	L2	L3	L4	L5	Total
AlexNet	10.15	123.25	43.98	65.95	43.99	287.34
2L-OPCNN	1.48		43.98	65.95	43.99	155.4
Five optical layers	1.48					1.48

By considering the information elaborated upon above, it is worth mentioning that because of parallel optical processing, both $\mathrm{T}_{\mathrm{operation1}}$ and $\mathrm{T}_{\mathrm{operation2}}$ nearly equal 45 ps. By considering $\mathrm{T}_{\mathrm{source}}$, $\mathrm{T}_{\mathrm{camera,}}$ and $\mathrm{T}_{\mathrm{transfer\cdot data,}}$ the 2L-Latency is estimated as 1.48 ms. Summarizing above discussion, considering CPU processor frequency of 3.8 GHz (Intel Core i7 8 core, Skylake-X microarchitecture), we calculated the computation delays of each convolutional layer for AlexNet and 2L-OPCNN assuming a grayscale input image with 227×227 pixels, as reported in Table V, whose details are provided in Section II of the supplementary material. It is worth mentioning that for instructions with various clock latencies [43], we considered the minimum number of clock to achieve the best execution time for the electrical design. In this manner, the achievable speedup by the optical network, compared to the electrical counterpart, can be higher in practice As an advantage of the optical processing, Table V concludes that the optical implementation of the first and the second convolutional layer results in speedup of 1.85 against AlexNet, by utilizing optical summation of channels' output for each convolutional kernel in the second layer structure, and so, concatenating two successive optical layers in 2L-OPCNN. On the other hand, as shown in Fig. 9, although the execution time of the electrical network increases by increasing the input image size, the processing time of the optical implementation does not depend on the input image size up to 3840 × 2160 pixels, which is the size of 4K UHD SLMs [44].

It is worth mentioning that although, as a case study, we investigated 2L-OPCNN with two successive optical layers, we can implement all five successive layers of AlexNet in the optical domain, considering optical units proposed in this paper. In this manner, we can achieve negligible latency of 1.48 ms, and speedup of 194.15 against the electrical AlexNet. However, it should be noted that optical implementation of five successive layers may face some challenges, like area consumption and alignment noises, which limit its feasibility as discussed in the next section.

It should be noted that in this work, the optical design is considered for the test procedure, while all training procedures are implemented electrically. To emphasize the importance of optical implementation of the forward inference in CNNs, optical processing of the large biological data sequences is explored as follows. As discussed in [45], classification of virus sequences (e.g., Coronaviruses, Dengue, HIV, Hepatitis B and C, and Influenza A), metagenomics data, and metabarcoding data can be performed by CNNs taking advantages of an appropriate image-based encoding method. It should be noted that single training procedure is carried out for each biological dataset while many test procedures are required to classify the input sequences. In this manner, optical implementation of the forward inference of the classifying CNNs would be greatly beneficial for biological datasets. For example, according to the influenza virus resource [46], 591280 sequences are collected till now which should be classified by a pretrained CNN. Moreover, classification of any new sequence, as NCBI database is continuously updated, do not require network retraining, while a pretrained CNN for the corresponding database can accomplish the classification task. Considering the proposed optical architecture, it is worth mentioning that the biological datasets can be encoded by the large images utilizing UHD SLMs of size 4K [44]. In this manner, for example, for classifying an input image with size of 2160×2160 pixels, the inference time of the first two layers of electrical AlexNet would be 13100 ms, while thanks to the optical implementation, the inference time of the first two layers of 2L-OPCNN is as small as 1.48 ms.

TABLE VI
AREA CONSUMPTION OF OPTICAL IMPLEMENTATION OF EACH CONVOLUTIONAL LAYERS OF ALEXNET

	L1-conv	L2-conv	L3-conv	L4-conv	L5-conv
Area	0.31 cm^2	79.85 cm^2	319.39 cm^2	479.08 cm^2	319.39 cm^2

V. SCALABILITY ANALYSIS

In term of experimental feasibility, implementing several optical layers is technically possible, but two physical concerns should be considered: I) the area consumption, and II) the alignment noises.

Area consumption depends on the number of kernels and the number of channels in each layer, since each kernel within a convolutional layer is optically implemented by a 4f system. Based on the total number of 4f systems and its cross sectional area, we can estimate the total area assuming all 4f systems in each layer are located adjacent to one another. Therefore, the total area of each convolutional layer can be estimated as follow:

$$Total\,area\,required = (number\ of\ kernels) * (number\ of\ channels) * (area\,of\,a\,single\,4f\ system) \quad (16)$$

To numerically calculate the total area, it is worth noting that we assumed the diameter of the lenses as 0.57 mm. Fig. 10 shows the side view of a 4f system whose dimension, as shown in green, is 0.57 mm. In this manner, each 4f system occupies an area of $(0.57\ \mathrm{mm})^2$, and so, the total area can be estimated based on the number of kernels and the number of channels of each layer, according to (16). Table VI reports the area consumption of an optical layer implementing the convolution operations within each layer of AlexNet. Correspondingly, Fig. 11 shows the area of each layer as the percentage of total area.

As shown in Fig. 11, area consumption by the first and the second layer is considerably less than those of the third, fourth, and fifth layers. It should be noted that the layers with the smaller number of masks are easier to implement, while arrays of metasurface lenses may be costly to fabricate for the layers with large number of kernels, although not practically impossible.

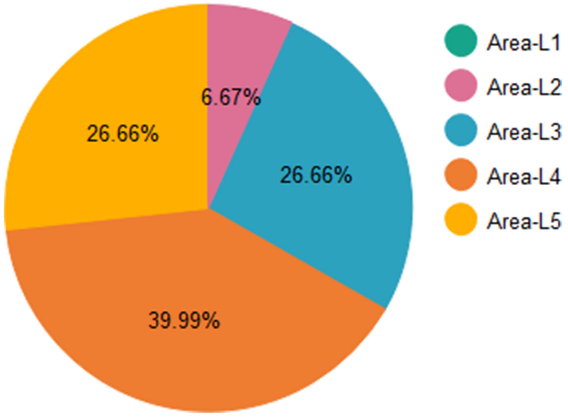

Fig. 11. Area percentage of each convolutional layer of total area.

To address the second physical concern, i.e., the alignment noises, we would like to emphasize that this issue is still a hot topic, while few researches [19], [47] have discussed it recently. Authors in [20] proposed an optronic CNN, however, due to the misalignment error in the optical setup, they enhanced the optronic network, as proposed in [47], and replaced the strided convolution layer by a spectral pooling layer. Also, they replaced the fully connected layer by a global average pooling (GAP) operation to increase the system's robustness to position variation of the input images. As another solution to this problem, we can improve the translation invariance property of the optical CNN, as briefly discussed in our previous work [19]. Currently, we are deeply exploring the idea that usage of a convolution-based pooling layer with appropriate trainable masks can considerably improve the translation invariance property of optical CNN, and so, the misalignment problem would be resolved. The detailed discussion of the corresponding simulation results would be published in near future. Finally, it is crystal clear that fabricating the filter masks and lenses in a monolithic manner (e.g., by semiconductor fab where the gaps between lenses and the lateral positioning is extremely precise) can considerably reduce the alignment noises.

VI. Power Analysis

As discussed in [19], the energy consumption for performing the optical convolution, nonlinearity, and pooling operations is negligible. Therefore, the main source of energy-consumption is the signal transduction which can be computed as follows [19], assuming $\sim 1\ \mu$W power of capturing each pixel at the detector side:

$$P_{optical} = \frac{n^2 \times n_{\ker nel}}{\eta \times t^p} \mu W, \tag{17}$$

where, n^2 is the total number of pixels per 4f correlator, p defines the number of optical elements through the path, n_{kernel} is the number of different kernels utilized in the convolutional layer, t is a fraction of incident power each optical element transmits, and finally, η is the source efficiency.

The total energy consumption of the electrical implementation is estimated as [19]:

$$P_{electronic} = \beta \times n^2 \times k^2 \times n_{\ker nel} \times P_{switching}, \tag{18}$$

where the constant coefficient β is determined by the executed architecture, k^2 is the kernels' size and $P_{\text{switching}}$ is the specific amount of energy consumed by each operation.

It should be noted that the energy consumption of both optical and electrical implementations scale, in a similar manner, with the number of pixels and the number of kernels. However, while the power consumption of the electronics parts depends on the kernel size, the power consumption of the optical implementation is independent form the kernel size. In view of these information elaborated upon, it is obvious that for the large kernel size, an optical implementation of the convolutional layers can significantly reduce the power consumption, compared to the electrical one.

VII. Conclusion

Recently, optical design has been proposed to improve performance of the deep neural networks. Although all-optical implementation of the CNNs has achieved many attentions, the recently proposed optical architectures for CNNs cannot fully utilize the tremendous capabilities of optical processing, due to the required electro-optical conversions in-between successive layers. To implement an all-optical multi-layer CNN, it is essential to optically implement all required operations, namely convolution, summation of channels' output for each convolutional kernel feeding the nonlinear unit, nonlinear activation function, and finally, pooling operation. In this paper, we explored a fully-optical design for implementing successive convolutional layers in an optical CNN. As a case study, we considered a CNN with two successive optical layers, named as 2L-OPCNN. The proposed architecture achieved 87.40%, 74.15%, and 99.34% accuracies for classifying images of Kaggle Cats and Dogs challenge, CIFAR-10, and MNIST datasets, respectively, which for Kaggle Cats and Dogs and MNIST datasets are almost the same classification accuracies provided by the electrical counterpart. Finally, it is worth noting that the 2L-OPCNN improved accuracies for various datasets, compared to the CNN utilizing a single optical layer. Also, a significant speedup were achieved by 2L-OPCNN against it electrical counterpart. There are still some issues with the proposed architecture which are considered as the future works. Although utilizing optical components provide considerable speedup in comparison with the electrical counterparts, they are more expensive and suffer from alignment noises. In this manner, an experimental feasibility study should consider the area consumption and address the translation invariance property of the ONNs to mitigate the alignment noises, as considered in our future works.

References

[1] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1207–1216, May 2016, doi: 10.1109/TMI.2016.2535865.

[2] S. Xu, X. Zou, B. Ma, J. Chen, L. Yu, and W. Zou, “Analog-to-digital conversion revolutionized by deep learning,” Oct. 2018, Accessed: Jun. 14, 2020. [Online]. Available: http://arxiv.org/abs/s1810.08906

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[4] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016 pp. 770–778.

[5] M. Miscuglio *et al.*, “Massively parallel amplitude-only Fourier neural network,” *Optica*, vol. 7, no. 12, pp. 1812–1819, 2020, doi: 10.1364/optica.408659.

[6] J. Chang, V. Sitzmann, X. Dun, and W. Heidrich, “Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification,” *Sci. Rep.*, vol. 8, no. 1, pp. 1–10, Dec. 2018, doi: 10.1038/s41598-018-30619-y.

[7] S. Colburn, Y. Chu, E. Shilzerman, and A. Majumdar, “Optical frontend for a convolutional neural network,” *Appl. Opt.*, vol. 58, no. 12, pp. 3179–3186, Apr. 2019, doi: 10.1364/AO.58.003179.

[8] L. De Marinis, M. Cococcioni, P. Castoldi, and N. Andriolli, “Photonic neural networks: A survey,” *IEEE Access*, vol. 7, pp. 175827–175841, 2019, doi: 10.1109/ACCESS.2019.2957245.

[9] F. Shokraneh, S. Geoffroy-Gagnon, M. S. Nezami, and O. Liboiron-Ladouceur, “A single layer neural network implemented by a 4×4 MZI-based optical processor,” *IEEE Photon. J.*, vol. 11, no. 6, Dec. 2019, Art. no. 4501612, doi: 10.1109/JPHOT.2019.2952562.

[10] M. Y.-S. Fang, S. Manipatruni, C. Wierzynski, A. Khosrowshahi, and M. R. DeWeese, “Design of optical neural networks with component imprecisions,” *Opt. Exp.*, vol. 27, no. 10, pp. 14009–14029, May 2019, doi: 10.1364/oe.27.014009.

[11] T. Yang, M. Chen, Y. Xiao, H. Xu, and P. Xu, “Research on 2F optical correlator based on neural network filter for recognizing large-angle rotation distortion target,” *IEEE Photon. J.*, vol. 12, no. 2, Apr. 2020, Art. no. 7800310, doi: 10.1109/JPHOT.2020.2970021.

[12] S. Xiang *et al.*, “A review: Photonics devices, architectures, and algorithms for optical neural computing,” *J. Semicond.*, vol. 42, no. 2, Feb. 2021, Art. no. 023105, doi: 10.1088/1674-4926/42/2/023105.

[13] S. Xiang *et al.*, “Computing primitive of fully VCSEL-based all-optical spiking neural network for supervised learning and pattern classification,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 6, pp. 2494–2505, Jun. 2021, doi: 10.1109/TNNLS.2020.3006263.

[14] A. A. Cruz-cabrera, M. Yang, G. Cui, E. C. Behrman, J. E. Steck, and S. R. Skinner, “Reinforcement and backpropagation training for an optical neural network using self-lensing effects,” *IEEE Trans. Neural Netw.*, vol. 11, no. 6, pp. 1450–1457, Nov. 2000, doi: 10.1109/72.883476.

[15] J. Liu *et al.*, “Research progress in optical neural networks: Theory, applications and developments,” *PhotoniX*, vol. 2, no. 1, pp. 1–39, Dec. 2021, doi: 10.1186/s43074-021-00026-0.

[16] X. Lin *et al.*, “All-optical machine learning using diffractive deep neural networks,” *Science*, vol. 361, no. 6406, pp. 1004–1008, Sep. 2018, doi: 10.1126/science.aat8084.

[17] J. Bueno *et al.*, “All-optical neural network with nonlinear activation functions,” *Optica*, vol. 6, no. 9, pp. 1132–1137, Sep. 2019, doi: 10.1364/optica.6.001132.

[18] Y. Shen *et al.*, “Deep learning with coherent nanophotonic circuits,” *Nature Photon.*, vol. 11, no. 7, pp. 441–446, Jun. 2017, doi: 10.1038/nphoton.2017.93.

[19] H. Sadeghzadeh, S. Koohi, and A. F. Paranj, “Free-space optical neural network based on optical nonlinearity and pooling operations,” *IEEE Access*, vol. 9, pp. 146533–146549, 2021, doi: 10.1109/access.2021.3123230.

[20] Z. Gu, Y. Gao, and X. Liu, “Optronic convolutional neural networks of multi-layers with different functions executed in optics for image classification,” *Opt. Exp.*, vol. 29, no. 4, pp. 5877–5889, 2021, doi: 10.1364/oe.415542.

[21] R. Hamerly, L. Bernstein, A. Sludds, M. Soljačić, and D. Englund, “Large-scale optical neural networks based on photoelectric multiplication,” *Phys. Rev. X*, vol. 9, no. 2, May 2019, Art. no. 021032, doi: 10.1103/PhysRevX.9.021032.

[22] K. Liao *et al.*, “All-optical computing based on convolutional neural networks,” *Opto-Electron. Adv.*, vol. 4, no. 11, pp. 1–9, 2021, doi: 10.29026/oea.2021.200060.

[23] Y. Jiang, W. Zhang, F. Yang, and Z. He, “Photonic convolution neural network based on interleaved time-wavelength modulation,” *J. Lightw. Technol.*, vol. 39, no. 14, pp. 4592–4600, Jul. 2021, doi: 10.1109/JLT.2021.3076070.

[24] Q. Wu *et al.*, “High speed and reconfigurable optronic neural network with digital nonlinear activation,” *Optik*, vol. 247, Dec. 2021, Art. no. 168043, doi: 10.1016/j.ijleo.2021.168043.

[25] A. Ryou *et al.*, “Free-space optical neural network based on thermal atomic nonlinearity,” *Photon. Res.*, vol. 9, no. 4, pp. B128–B134, Apr. 2021, doi: 10.1364/prj.415964.

[26] Z. Hu, M. Miscuglio, J. George, Y. Alkabani, T. El Gazhawi, and V. J. Sorger, “Highly-parallel optical Fourier intensity convolution filter for image classification,” in *Proc. Opt. InfoBase Conf. Paper*, 2019, pp. 100–102, doi: 10.1364/FIO.2019.JW4A.101.

[27] T. Harasthy, Ľ. Ovseník, and J. Turán, “Current summary of the practical using of optical correlators,” *Acta Electrotechnica Inform.*, vol. 12, no. 4, pp. 30–38, Jan. 2012, doi: 10.2478/v10198-012-0042-2.

[28] Q. Wu *et al.*, “Multi-layer optical Fourier neural network based on the convolution theorem,” *AIP Adv.*, vol. 11, no. 5, 2021, Art. no. 055012, doi: 10.1063/5.0055446.

[29] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, “Striving for simplicity: The all convolutional net,” in *Proc. 3rd Int. Conf. Learn. Represent Work. Track*, 2015, pp. 1–14.

[30] S. Akbari Rokn Abadi, N. Hashemi Dijujin, and S. Koohi, “Optical pattern generator for efficient bio-data encoding in a photonic sequence comparison architecture,” *PLoS One*, vol. 16, no. 1, Jan. 2021, Art. no. e0245095, doi: 10.1371/journal.pone.0245095.

[31] J. R. Ong, C. C. Ooi, T. Y. L. Ang, S. T. Lim, and C. E. Png, “Photonic convolutional neural networks using integrated diffractive optics,” *IEEE J. Sel. Topics Quantum Electron.*, vol. 26, no. 5, Sep./Oct. 2020, Art. no. 7702108, doi: 10.1109/JSTQE.2020.2982990.

[32] X. Guo, T. D. Barrett, Z. M. Wang, and A. I. Lvovsky, “Backpropagation through nonlinear units for the all-optical training of neural networks,” *Photon. Res.*, vol. 9, no. 3, pp. B71–B80, Mar. 2021, doi: 10.1364/prj.411104.

[33] A. Vander Lugt, “Signal detection by complex spatial filtering,” *IEEE Trans. Inf. Theory*, vol. 10, no. 2, pp. 139–145, Apr. 1964, doi: 10.1109/TIT.1964.1053650.

[34] X. Deng, X. Zhang, Y. Wang, Y. Song, S. Liu, and C. Li, “Intensity threshold in the conversion from reverse saturable absorption to saturable absorption and its application in optical limiting,” *Opt. Commun.*, vol. 168, no. 1–4, pp. 207–212, Sep. 1999, doi: 10.1016/S0030-4018(99)00297-7.

[35] R. Ayachi, M. Afif, Y. Said, and M. Atri, “Strided convolution instead of max pooling for memory efficiency of convolutional neural networks,” *Smart Innov. Syst. Technol.*, vol. 146, pp. 234–243, 2020, doi: 10.1007/978-3-030-21005-2_23.

[36] S. Ngcobo, I. Litvin, L. Burger, and A. Forbes, “A digital laser for on-demand laser modes,” *Nature Commun.*, vol. 4, no. 1, pp. 1–6, Oct. 2013, doi: 10.1038/ncomms3289.

[37] F. Aieta *et al.*, “Aberration-free ultrathin flat lenses and axicons at telecom wavelengths based on plasmonic metasurfaces,” *Nano Lett.*, vol. 12, no. 9, pp. 4932–4936, 2012, doi: 10.1021/nl302516v.

[38] Vapor Reference Cells, Accessed: Jun. 8, 2022. [Online]. Available: https://www.thorlabs.com/newgrouppage9.cfm?objectgroup_id=1470

[39] Dogs vs. Cats | Kaggle, Accessed: Jun. 8, 2022. [Online]. Available: https://www.kaggle.com/c/dogs-vs-cats/data

[40] A. Krizhevsky, “Learning multiple layers of features from tiny images,” M.S. thesis, Dept. Comput. Sci, Univ. Toronto, Toronto, ON, Canada, 2009.

[41] Y. LeCun, C. Cortes, and C. Burges, “MNIST handwritten digit database,” Accessed: Mar. 1, 2021, [Online]. Available: http://yann.lecun.com/exdb/mnist/

[42] HS7 – High Speed Imaging, Accessed: Apr. 13, 2021. [Online]. Available: https://hsi.ca/product/hs7/

[43] A. Fog, “Introduction 4. Instruction tables,” 1996, [Online]. Available: www.agner.org/optimize/testp.zip

[44] Spatial Light Modulators, Accessed: May 23, 2022. [Online]. Available:https://www.thorlabs.com/newgrouppage9.cfm?objectgroup_id=10378

[45] S. Akbari Rokn Abadi, A. Mohammadi, and S. Koohi, “WalkIm: Compact image-based encoding for high-performance classification of biological sequences using simple tuning-free CNNs,” *PLoS One*, vol. 17, no. 4, 2022, Art. no. e0267106, doi: 10.1371/journal.pone.0267106.

[46] Influenza virus database - NCBI, Accessed: May 18, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/genomes/FLU/Database/nph-select.cgi#mainform()

[47] Z. Gu, Y. Gao, and X. Liu, “Position-robust optronic convolutional neural networks dealing with images position variation,” *Opt. Commun.*, vol. 505, 2022, Art. no. 127505, doi: 10.1016/j.optcom.2021.127505.