Reprogrammable Electro-Optic Nonlinear Activation Functions for Optical Neural Networks

Ian A. D. Williamson ⓘ, *Member, IEEE*, Tyler W. Hughes ⓘ, Momchil Minkov ⓘ, Ben Bartlett ⓘ, Sunil Pai ⓘ, and Shanhui Fan, *Fellow, IEEE*

(Invited Paper)

Abstract—We introduce an electro-optic hardware platform for nonlinear activation functions in optical neural networks. The optical-to-optical nonlinearity operates by converting a small portion of the input optical signal into an analog electric signal, which is used to intensity -modulate the original optical signal with no reduction in processing speed. Our scheme allows for complete nonlinear ON–OFF contrast in transmission at relatively low optical power thresholds and eliminates the requirement of having additional optical sources between each of the layers of the network Moreover, the activation function is reconfigurable via electrical bias, allowing it to be programmed or trained to synthesize a variety of nonlinear responses. Using numerical simulations, we demonstrate that this activation function significantly improves the expressiveness of optical neural networks, allowing them to perform well on two benchmark machine learning tasks: learning a multi-input exclusive-OR (XOR) logic function and classification of images of handwritten numbers from the MNIST dataset. The addition of the nonlinear activation function improves test accuracy on the MNIST task from 85% to 94%.

Index Terms—Optical neural networks, feedforward neural networks, neuromorphic computing, machine learning, electro-optic modulators, photodetectors, nonlinear optics, intensity modulation, phase modulation.

I. INTRODUCTION

IN RECENT years, there has been significant interest in alternative computing platforms specialized for high performance and efficiency on machine learning tasks. For example, graphical processing units (GPUs) have demonstrated peak performance with trillions of floating point operations per second (TFLOPS) when performing matrix multiplication, which is several orders of magnitude larger than general-purpose digital processors such as CPUs [1]. Moreover, analog computing has been explored for achieving high performance because it is not limited by the bottlenecks of sequential instruction execution and memory access [2]–[6].

Optical hardware platforms are particularly appealing for computing and signal processing due to their ultra-large signal bandwidths, low latencies, and reconfigurability [7]–[9]. They have also gathered significant interest in machine learning applications, such as artificial neural networks (ANNs). Nearly three decades ago, the first optical neural networks (ONNs) were proposed based on free-space optical lens and holography setups [10], [11]. More recently, ONNs have been implemented in chip-integrated photonic platforms [12] using programmable waveguide interferometer meshes which perform matrix-vector multiplications [13]. In theory, the performance of such systems is competitive with digital computing platforms because they may perform matrix-vector multiplications in constant time with respect to the matrix dimension. In contrast, matrix-vector multiplication has a quadratic time complexity on a digital processor. Other approaches to performing matrix-vector multiplications in chip-integrated ONNs, such as microring weight banks and photodiodes, have also been proposed [14].

Nonlinear activation functions play a key role in ANNs by enabling them to learn complex mappings between their inputs and outputs. Whereas digital processors have the expressiveness to trivially apply nonlinearities such as the widely-used sigmoid, ReLU, and tanh functions, the realization of nonlinearities in optical hardware platforms is more challenging. One reason for this is that optical nonlinearities are relatively weak, necessitating a combination of large interaction lengths and high signal powers, which impose lower bounds on the physical footprint and the energy consumption, respectively. Although it is possible to resonantly enhance optical nonlinearities, this comes with an unavoidable trade-off in reducing the operating bandwidth, thereby limiting the information processing capacity of an ONN. Additionally, maintaining uniform resonant responses across many elements of an optical circuit necessitates additional control circuitry for calibrating each element [15].

A more fundamental limitation of optical nonlinearities is that their responses tend to be fixed during device fabrication. This limited tunability of the nonlinear optical response prevents an ONN from being reprogrammed to realize different forms of nonlinear activation functions, which may be important for

tailoring ONNs for different machine learning tasks. Similarly, a fixed nonlinear response may also limit the performance of very deep ONNs with many layers of activation functions since the optical signal power drops below the activation threshold, where nonlinearity is strongest, in later layers due to loss in previous layers. For example, with optical saturable absorption from 2D materials in waveguides, the activation threshold is on the order of 1–10 mW [16]–[18], meaning that the strength of the nonlinearity in each subsequent layer will be successively weaker as the transmitted power falls below the threshold.

In light of these challenges, the ONN demonstrated in Ref. [12] implemented its activation functions by detecting each optical signal, feeding them through a conventional digital computer to apply the nonlinearity, and then modulating new optical signals for the subsequent layer. Although this approach benefits from the flexibility of digital signal processing, conventional processors have a limited number of input and output channels, which make it challenging to scale this approach to very large matrix dimensions, which corresponds to a large number of optical inputs. Moreover, digitally applied nonlinearities add latency from the analog-to-digital conversion process and constrain the computational speed of the neural network to the same GHz-scale clock rates which ONNs seek to overcome. Thus, a hardware nonlinear optical activation, which doesn't require repeated bidirectional optical-electronic signal conversion, is of fundamental interest for making integrated ONNs a viable machine learning platform.

In this article, we propose an electro-optic architecture for synthesizing optical-to-optical nonlinearities which alleviates the issues discussed above. Our architecture features complete *on-off* contrast in signal transmission, a variety of nonlinear response curves, and a low activation threshold. Rather than using traditional optical nonlinearities, our scheme operates by measuring a small portion of the incoming optical signal power and using electro-optic modulators to modulate the original optical signal, without any reduction in operating bandwidth or computational speed. Additionally, our scheme allows for the possibility of performing additional nonlinear transformations on the signal using analog electrical components. Related electro-optical architectures for generating optical nonlinearities have been previously considered [19]–[21]. In this work, we focus on the application of our architecture as an element-wise activation in a feedforward ONN, but the synthesis of low-threshold optical nonlinearities could be of broader interest to optical computing and information processing.

The remainder of this paper is organized as follows. First, we review the basic operating principles of ANNs and their integrated optical implementations in waveguide interferometer meshes. We then introduce our electro-optical activation function architecture, showing that it can be reprogrammed to synthesize a variety of nonlinear responses. Next, we discuss the performance of an ONN using this architecture by analyzing the scaling of power consumption, latency, processing speed, and footprint. We then draw an analogy between our proposed activation function and the optical Kerr effect. Finally, using numerical simulations, we demonstrate that our architecture
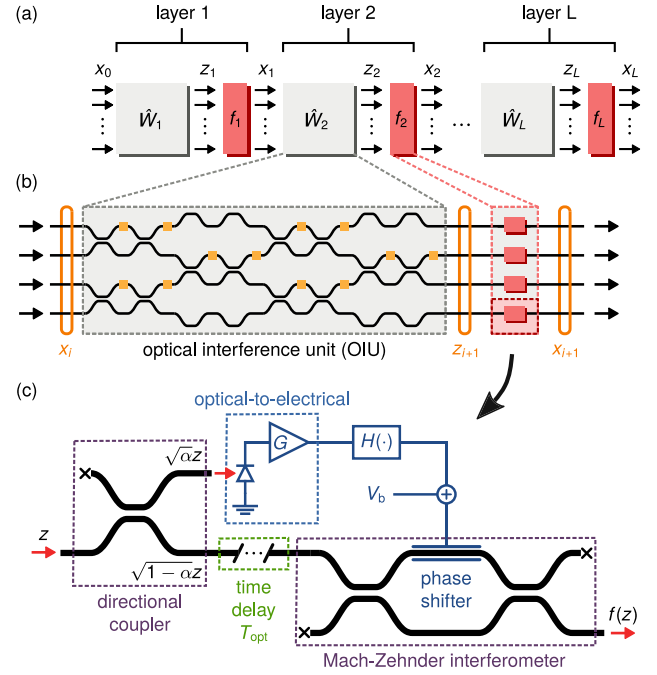

Fig. 1. (a) Block diagram of a feedforward neural network of L layers. Each layer consists of a $\hat{W}_i$ block representing a linear matrix which multiplies vector inputs x_{i-1}. The f_i block in each layer represents an element-wise nonlinear activation function operating on vectors z_i to produce outputs x_i. (b) Schematic of the optical interferometer mesh implementation of a single layer of the feedforward neural network. (c) Schematic of the proposed optical-to-optical activation function which achieves a nonlinear response by converting a small portion of the optical input, z into an electrical signal, and then intensity modulating the remaining portion of the original optical signal as it passes through an interferometer.

leads to improved performance on two different machine learning tasks: (1) learning an N-input exclusive OR (XOR) logic function; (2) classifying images of handwritten numbers from the MNIST dataset.

II. FEEDFORWARD OPTICAL NEURAL NETWORKS

In this section, we briefly review the basics of feedforward artificial neural networks (ANNs) and describe their implementation in a reconfigurable optical circuit, as proposed in Ref. [12]. As outlined in Fig. 1(a), an ANN is a function which accepts an input vector, x_0 and returns an output vector, x_L. This is accomplished in a layer-by-layer fashion, with each layer consisting of a linear matrix-vector multiplication followed by the application of an element-wise nonlinear function, or *activation*, on the result. For a layer with index i, containing a weight matrix $\hat{W}_i$ and activation function $f_i(\cdot)$, its operation is described mathematically as

$$x_i = f_i\left(\hat{W}_i \cdot x_{i-1}\right) \qquad (1)$$

for i from 1 to L.

Before they are able to perform a given machine learning task, ANNs must be trained. The training process is typically accomplished by minimizing the prediction error of the ANN on a set of training examples, which come in the form of input and

target output pairs. For a given ANN, a loss function is defined to quantify the difference between the target output and output predicted by the network. During training, this loss function is minimized with respect to tunable degrees of freedom, namely the elements of the weight matrix $\hat{W}_i$ within each layer. In general, although less common, it is also possible to train the parameters of the activation functions [22].

Optical hardware implementations of ANNs have been proposed in various forms over the past few decades. In this work, we focus on a recent demonstration in which the linear operations are implemented using an integrated optical circuit [12]. In this scheme, the information being processed by the network, x_i, is encoded into the modal amplitudes of the waveguides feeding the device and the matrix-vector multiplications are accomplished using meshes of integrated optical interferometers. In this case, training the network requires finding the optimal settings for the integrated optical phase shifters controlling the inteferometers, which may be found using an analytical model of the chip, or using *in-situ* backpropagation techniques [23].

In the next section, we present an approach for realizing the activation function, $f_i(\cdot)$, on-chip with a hybrid electro-optic circuit feeding an inteferometer. In Fig. 1(b), we show how this activation scheme fits into a single layer of an ONN and show the specific form of the activation in Fig. 1(c). We also give the specific mathematical form of this activation and analyze its performance in practical operation.

III. NONLINEAR ACTIVATION FUNCTION ARCHITECTURE

In this section, we describe our proposed nonlinear activation function architecture for optical neural networks, which implements an optical-to-optical nonlinearity by converting a small portion of the optical input power into an electrical voltage. The remaining portion of the original optical signal is phase- and amplitude-modulated by this voltage as it passes through an interferometer. For an input signal with amplitude z, the resulting nonlinear optical activation function, $f(z)$, is a result of the responses of the interferometer under modulation as well as the components in the electrical signal pathway.

A schematic of the architecture is shown in Fig. 1(c), where black and blue lines represent optical waveguides and electrical signal pathways, respectively. The input signal first enters a directional coupler which routes a portion, α, of the input optical power to a photodetector. The photodetector is the first element of an optical-to-electrical conversion circuit, which is a standard component of high-speed optical receivers for converting an optical intensity into a voltage. In this work, we assume a normalization of the optical signal such that the total power in the input signal is given by $|z|^2$. The optical-to-electrical conversion process consists of the photodetector producing an electrical current, $I_{pd} = \mathfrak{R} \cdot \alpha|z|^2$, where $\mathfrak{R}$ is the photodetector responsivity, and a transimpedance amplifying stage, characterized by a gain G, converting this current into a voltage $V_G = G \cdot \mathfrak{R} \cdot \alpha|z|^2$. The output voltage of the optical-to-electrical conversion circuit then passes through a nonlinear signal conditioner with a transfer function, $H(\cdot)$. This component allows for the application of additional nonlinear functions to transform the voltage signal.

Finally, the conditioned voltage signal, $H(V_G)$ is combined with a static bias voltage, V_b to induce a phase shift of

$$\Delta\phi = \frac{\pi}{V_\pi}\left[V_b + H\left(G\mathfrak{R}\alpha|z|^2\right)\right] \qquad (2)$$

for the optical signal routed through the lower port of the directional coupler. The parameter V_π represents the voltage required to induce a phase shift of π in the phase modulator. This phase shift, defined by Eq. 2, is a nonlinear self-phase modulation because it depends on the input signal intensity.

An optical delay line between the directional coupler and the Mach-Zehnder interferometer (MZI) is used to match the signal propagation delays in the optical and electrical pathways. This ensures that the nonlinear self-phase modulation defined by Eq. (2) is applied at the same time that the optical signal which generated it passes through the phase modulator. For the circuit shown in Fig. 1(c), the optical delay is $\tau_{opt} = \tau_{oe} + \tau_{nl} + \tau_{rc}$, accounting for the contributions from the group delay of the optical-to-electrical conversion stage (τ_{oe}), the delay associated with the nonlinear signal conditioner (τ_{nl}), and the RC time constant of the phase modulator (τ_{rc}).

The nonlinear self-phase modulation achieved by the electric circuit is converted into a nonlinear amplitude response by the MZI, which has a transmission depending on $\Delta\phi$ as

$$t_{MZI} = j\exp\left(-j\frac{\Delta\phi}{2}\right)\cos\left(\frac{\Delta\phi}{2}\right). \qquad (3)$$

Depending on the configuration of the bias, V_b, a larger input optical signal amplitude causes either more or less power to be diverted away from the output port, resulting in a nonlinear self-intensity modulation. Combining the expression for the nonlinear self-phase modulation, given by Eq. 2, with the MZI transmission, given by Eq. 3, the mathematical form of the activation function can be written explicitly as

$$f(z) = j\sqrt{1-\alpha}\exp\left(-j\frac{1}{2}\left[\phi_b + \pi\frac{H\left(G\mathfrak{R}\alpha|z|^2\right)}{V_\pi}\right]\right)$$
$$\cdot \cos\left(\frac{1}{2}\left[\phi_b + \pi\frac{H\left(G\mathfrak{R}\alpha|z|^2\right)}{V_\pi}\right]\right)z, \qquad (4)$$

where the contribution to the phase shift from the bias voltage is

$$\phi_b = \pi\frac{V_b}{V_\pi}. \qquad (5)$$

For the remainder of this work, we focus on the case where no nonlinear signal conditioning is applied to the electrical signal pathway, i.e. $H(V_G) = V_G$. However, even with this simplification the activation function still exhibits a highly nonlinear response. We also neglect saturating effects in the OE conversion stage which can occur in either the photodetector or the amplifier. However, in practice, the nonlinear optical-to-optical transfer function could take advantage of these saturating effects.

With the above simplifications, a more compact expression for the activation function response is

$$f(z) = j\sqrt{1-\alpha}\exp\left(-j\left[\frac{g_\phi\,|z|^2}{2} + \frac{\phi_b}{2}\right]\right)$$
$$\cdot \cos\left(\frac{g_\phi\,|z|^2}{2} + \frac{\phi_b}{2}\right)z, \qquad (6)$$

where the phase gain parameter is defined as

$$g_\phi = \pi\frac{\alpha G\Re}{V_\pi}. \qquad (7)$$

Equation 7 indicates that the amount of phase shift per unit input signal power can be increased via the gain and photodiode responsivity, or by converting a larger fraction of the optical power to the electrical domain. However, tapping out a larger fraction optical power also results in a larger linear loss, which is undesirable.

The electrical biasing of the activation phase shifter, represented by V_b, is an important degree of freedom for determining its nonlinear response. We consider a representative selection, consisting of four different responses, in Fig. 2. The left column of Fig. 2 plots the output signal amplitude as a function of the input signal amplitude i.e. $|f(z)|$ in Eq. 6, while the right column plots the transmission coefficient i.e. $|f(z)|^2/|z|^2$, a quantity which is more commonly used in optics than machine learning. The first two rows of Fig. 2, corresponding to $\phi_b = 1.0\pi$ and 0.85π, exhibit a response which is comparable to the ReLU activation function: transmission is low for small input values and high for large input values. For the bias of $\phi_b = 0.85\pi$, transmission at low input values is slightly increased with respect to the response where $\phi_b = 1.00\pi$. Unlike the ideal ReLU response, the activation at $\phi_b = 0.85\pi$ is not entirely monotonic because transmission first goes to zero before increasing. On the other hand, the responses shown in the bottom two rows of Fig. 2, corresponding to $\phi_b = 0.0\pi$ and 0.50π, are quite different. These configurations demonstrate a saturating response in which the output is suppressed for higher input values but enhanced for lower input values. For all of the responses shown in Fig. 2, we have assumed $\alpha = 0.1$ which limits the maximum transmission to $1 - \alpha = 0.9$.

A benefit of having electrical control over the activation response is that, in principle, its electrical bias can be connected to the same control circuitry which programs the linear interferometer meshes. In doing so, a single ONN hardware unit can then be reprogrammed to synthesize many different activation function responses. This opens up the possibility of heuristically selecting an activation function response, or directly optimizing the the activation bias using a training algorithm. This realization of a flexible optical-to-optical nonlinearity can allow ONNs to be applied to much broader classes of machine learning tasks.

We note that Fig. 2 shows only the amplitude response of the activation function. In fact, all of these responses also introduce a nonlinear self-phase modulation to the output signal. If desired, this nonlinear self-phase modulation can be suppressed using a push-pull interferometer configuration in which the generated
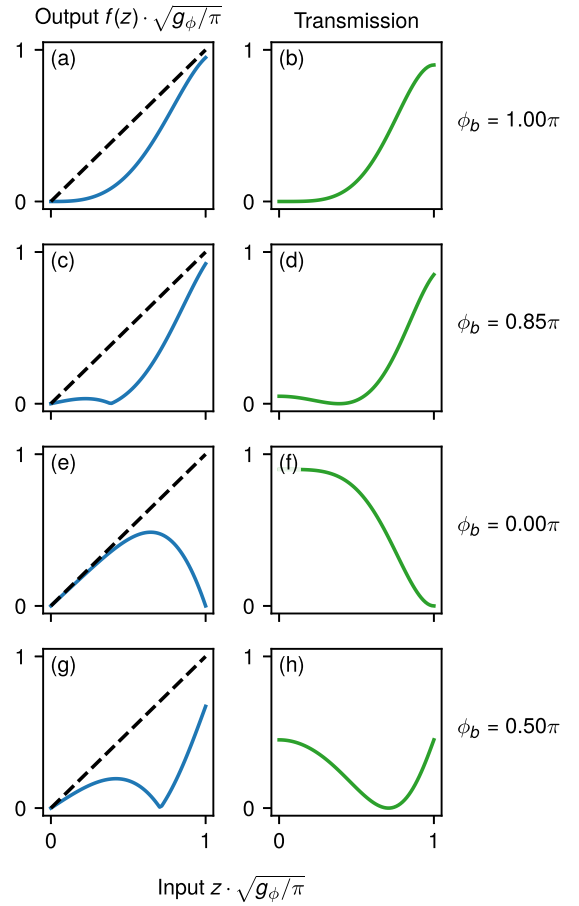

Fig. 2. Activation function output amplitude (blue lines) and activation function transmission (green lines) as a function of input signal amplitude. The input and output are normalized to the phase gain parameter, g_ϕ. Panel pairs (a), (b) and (c), (d) correspond to a ReLU-like response, with a suppressed transmission for inputs with small amplitude and high transmission for inputs with large amplitude. Panel pairs (e), (f) and (g), (h) correspond to a clipped response, with high transmission for inputs with small amplitude and reduced transmission for inputs with larger amplitude.

phase shift, $\Delta\phi$, is divided and applied with opposite sign to the top and bottom arms.

IV. PERFORMANCE AND SCALABILITY

In this section, we discuss the performance of an integrated ONN which uses meshes of integrated optical interferometers to perform matrix-vector multiplications and the electro-optic activation function, as shown in Fig. 1(b),(c). Here, we focus on characterizing how the power consumption, computational latency, physical footprint, and computational speed of the ONN scale with respect to the number of network layers, L and the dimension of the input vector, N, assuming square matrices. The system parameters used for this analysis are summarized in Table I and the figures of merit are summarized in Table II.

A. Power Consumption

The power consumption of the ONN, as shown in Fig. 1(b), consists of contributions from (1) the programmable phase

Parameter	Value
Modulator and detector rate	10 GHz
Photodetector responsivity ($\mathfrak{R}$)	1 A/W
Optical-to-electrical circuit power consumption	100 mW
Optical-to-electrical circuit group delay (τ_{eo})	100 ps
Phase modulator RC delay (τ_{rc})	20 ps
Mesh MZI length (D_{MZI})	100 μm
Mesh MZI height (H_{MZI})	60 μm
Waveguide effective index (n_{eff})	3.5

shifters inside the interferometer mesh, (2) the optical source supplying the input vectors, x_0, and (3) the active components of the activation function such as the amplifier and photodetector. In principle, the contribution from (1) can be made negligible by using phase change materials or ultra-low power MEMS phase shifters. Therefore, in this section we focus only on contributions (2) and (3) which pertain to the activation function.

To quantify the power consumption, we first consider the minimum input optical power to a *single* activation that triggers a nonlinear response. We refer to this as the activation function threshold, which is mathematically defined as

$$P_{th} = \frac{\Delta\phi|_{\delta T=0.5}}{g_\phi} = \frac{V_\pi}{\pi\alpha G\mathfrak{R}} \cdot \Delta\phi|_{\delta T=0.5}, \qquad (8)$$

where $\Delta\phi|_{\delta T=0.5}$ is the is phase shift necessary to generate a 50% change in the power transmission with respect to the transmission with null input for a given ϕ_b. This threshold corresponds to $z\sqrt{g_\phi/\pi} = 0.73$ in Fig. 2(b), to $z\sqrt{g_\phi/\pi} = 0.85$ in Fig. 2(d), to $z\sqrt{g_\phi/\pi} = 0.73$ in Fig. 2(f), and to $z\sqrt{g_\phi/\pi} = 0.70$ in Fig. 2(h). In general, a lower activation threshold will result in a lower optical power required at the ONN input, $|x_0|^2$. According to Eq. (8), the activation threshold can be reduced via a small V_π and a large optical-to-electrical conversion gain, $G\mathfrak{R} \sim 1.0$ V/mW. The relationship between G and V_π for activation thresholds of 0.1 mW, 1.0 mW, and 10.0 mW is shown in Fig. 3 for a fixed $\mathfrak{R} = 1$ A/W. Additionally, in Fig. 3 we conservatively assume $\phi_b = \pi$ which has the highest threshold of the activation function biases shown in Fig. 2.

If we take the lowest activation threshold of 0.1 mW in Fig. 3, the optical source to the ONN would then need to supply $N \cdot 0.1$ mW of optical power. The power consumption of integrated optical receiver amplifiers varies considerably, ranging from as low as 10 mW to as high as 150 mW [24]–[26], depending on a variety of factors which are beyond the scope of this article. Therefore, a conservative estimate of the power consumption from the optical-to-electrical conversion circuits in all activations is $L \cdot N \cdot 100$ mW. For an ONN with $N = 100$, the power consumption per layer from the activation function would be 10 W and would require a total optical input power of $N \cdot P_{th} = 100 \cdot 0.1$ mW $= 10$ mW. Thus, the total power consumption of the ONN is dominated by the activation function electronics.

B. Latency

For the feedforward neural network architecture shown in Fig. 1(a), the latency is defined by the elapsed time between supplying an input vector, x_0 and detecting its corresponding prediction vector, x_L. In an integrated ONN, as implemented in Fig. 1(b), this delay is simply the travel time for an optical pulse through all L-layers. Following Ref. [12], the propagation distance in a square interferometer mesh is $D_W = N \cdot D_{MZI}$, where D_{MZI} is the length of each MZI within the mesh. In the nonlinear activation layer, the propagation length will be dominated by the delay line required to match the optical and electrical delays, and is given by

$$D_f = (\tau_{oe} + \tau_{nl} + \tau_{rc}) \cdot v_g, \qquad (9)$$

where the group velocity $v_g = c_0/n_{eff}$ is the speed of optical pulses in the waveguide. Therefore,

$$\text{latency} = \underbrace{L \cdot N \cdot D_{MZI} \cdot v_g^{-1}}_{\text{Interferometer mesh}} + \underbrace{L \cdot (\tau_{oe} + \tau_{nl} + \tau_{rc})}_{\text{Activation function}}. \qquad (10)$$

Equation 10 indicates that the latency contribution from the interferometer mesh scales with the product LN, which is the same scaling as predicted in Ref. [12]. On the other hand, the activation function adds to the latency independently of N because each activation circuit is applied in parallel to all N-vector elements.

For concreteness, we assume $D_{MZI} = 100 \, \mu$m and $n_{eff} = 3.5$. Following our assumption in the previous section of using no nonlinear electrical signal conditioner in the activation function, $\tau_{nl} = 0$ ps. Typical group delays for integrated transimpedance amplifiers used in optical receivers can range from $\tau_{oe} \approx 10$ to 100 ps. Moreover, assuming an RC-limited phase modulator speed of 50 GHz yields $\tau_{rc} \approx 20$ ps. Therefore, if we assume a conservative value of $\tau_{oe} = 100$ ps, a network dimension of $N \approx 100$ would have a latency of 237 ps per layer, with equal contributions from the mesh and the activation function. For a ten layer network ($L = 10$) the total latency would be approximately 2.4 ns, still orders of magnitude lower than the latency typically associated with GPUs.

C. Physical Footprint

The physical footprint of the ONN consists of the space taken up by both the linear interferometer mesh and the optical and electrical components of the activation function. Neglecting the electrical control lines for tuning each MZI, the total footprint of the ONN is

$$A = \underbrace{L \cdot N^2 \cdot A_{MZI}}_{\text{Interferometer mesh}} + \underbrace{L \cdot N \cdot A_f}_{\text{Activation function}}, \qquad (11)$$

where $A_{MZI} = D_{MZI} \cdot H_{MZI}$ is the area of a single MZI element in the mesh and $A_f = D_f \cdot H_f$ is the area of a single activation function.

In the direction of propagation, D_f is dominated by the waveguide optical delay line required to match the delay of the electrical signal pathway. Based on the previous discussion of the activation function's latency, $\tau_{opt} = 120$ ps corresponds to a total waveguide length of $D_f \approx 1$ cm. For simplicity, we assume

TABLE II
SUMMARY OF PER-LAYER OPTICAL NEURAL NETWORK PERFORMANCE USING THE ELECTRO-OPTIC ACTIVATION FUNCTION

	Scaling		Per-layer figures of merit		
	Mesh	Activation	$N = 4$	$N = 10$	$N = 100$
Power consumption*		LN	0.4 W	1 W	10 W
Latency	LN	L	125 ps	132 ps	237 ps
Footprint	LN^2	LN	2.5 mm^2	6.6 mm^2	120.0 mm^2
Speed	LN^2		1.6×10^{11} MAC/s	1×10^{12} MAC/s	1×10^{14} MAC/s
Efficiency*		N^{-1}	2.5 pJ/MAC	1 pJ/MAC	100 fJ/MAC

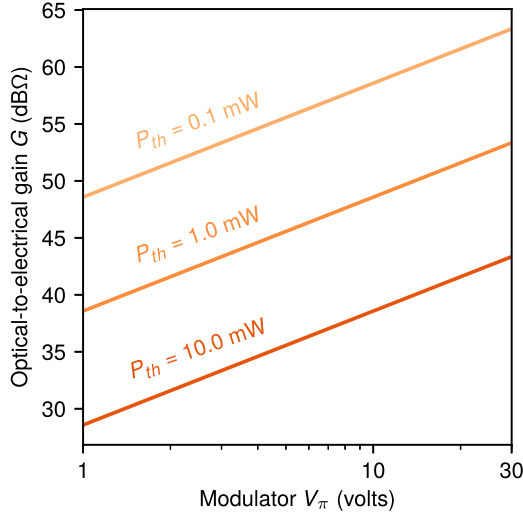*Assuming no power consumption in the interferometer mesh phase shifters.

Fig. 3. Contours of constant activation threshold as a function of the optical-to-electrical gain and the modulator V_π of the activation function shown in Fig. 1(c) with a photodetector responsivity $\Re = 1.0$ A/W.

this delay is achieved using a straight waveguide, which results in a large footprint but with optical losses that can be very low. For example, in silicon waveguides losses below 0.5 dB/cm have been experimentally demonstrated [27]. In principle, incorporating waveguide bends or resonant optical elements could significantly reduce the activation function's footprint. For example, coupled micro ring arrays have experimentally achieved group delays of 135 ps over a bandwidth of 10 GHz in a 0.03 mm × 0.25 mm footprint [28].

Transverse to the direction of propagation, the activation function footprint will be dominated by the electronic components of the optical-to-electrical conversion circuit. In principle, compact waveguide photodetectors and modulators can be utilized. However, the components of the transimpedance amplifier may be challenging to integrate in the area available between neighboring output waveguides of the interferometer mesh. One possibility towards achieving a fully integrated opto-electronic ONN would be to use so-called *amplifier-free* optical receivers [26], where ultra-low capacitance detectors provide high-speed opto-electronic conversion. Similarly to the experimental demonstration in Ref. [29], the amplifier-free receiver could be integrated directly with a high efficiency (e.g. effectively a low V_π) electro-optic modulator. Compact electro-absorption modulators could also be utilized. In addition to achieving a compact footprint,

operating without an amplifier would also result in an order of magnitude reduction in both power consumption and latency, with the later reducing the required length of the optical delay line and thus the footprint.

For the purposes of our analysis, we assume no integration of the electronic transimpedance amplifier and, therefore, that the on-chip components of the activation function fit within the height of each interferometer mesh row, $D_f \leq D_{\mathrm{MZI}} = 60$ μm. Under this assumption and following the scaling in Eq. 11, the total footprint of a single ONN layer of dimension $N = 10$ would be 11.0 mm × 0.6 mm. Interestingly, following the latency discussion in the previous section, a single ONN layer of dimension $N = 100$ would have a footprint of 20.0 mm × 6.0 mm, with equal contribution from the activation function and from the mesh.

D. Speed

The speed, or computational capacity, of the ONN, as shown in Fig. 1(a), is determined by the number of input vectors, x_0 that can be processed per unit time. Here, we argue that although our activation function is not fully optical, it results in no speed degradation compared to a linear ONN consisting of only interferometer meshes.

The reason for this is that a fully integrated ONN would also include high-speed modulators and detectors on-chip to perform fast modulation and detection of sequences of x_0 vectors and x_L vectors, respectively. We therefore argue that the same high-speed detector and modulator elements could also be integrated between the linear network layers to provide the optical-electrical and electrical-optical transduction for the activation function. State of the art integrated transimpedance amplifiers can already operate at speeds comparable to the optical modulator and detector rates, which are on the order of 50–100 GHz [24], [30], and thus would not be a limiting factor in the speed of our architecture.

To perform a matrix-vector multiplication on a conventional CPU requires N^2 multiply-accumulate (MAC) operations, each consisting of a single multiplication and a single addition. Therefore, assuming a photodetector and modulator rate of 10 GHz means that an ONN can effectively perform $N^2 \cdot L \cdot 10^{10}$ MAC/sec. This means that one layer of an ONN with dimension $N = 10$ would effectively perform 10^{12} MAC/sec. Increasing the input dimension to $N = 100$ would then scale the performance of the ONN to 10^{14} MAC/sec per layer. This is two 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). Because the power consumption of the ONN scales as LN (assuming passive phase shifters in the mesh) and the speed scales as LN^2, the energy per operation is minimized for large N (Table II). Thus, for large ONNs the power consumption associated with the electro-optic conversion in the activation function can be amortized over the parallelized operation of the linear mesh.

We note that the activation function circuit shown in Fig. 1(c) can be modified to remove the matched optical delay line by using very long optical pulses. This modification may be advantageous for reducing the footprint of the activation and would result in $\tau_{\text{opt}} \ll \tau_{\text{ele}}$. However, this results in a reduction of the ONN speed, which would then be limited by the combined activation delay of all L nonlinear layers in the network, $\sim (L \cdot \tau_{\text{ele}})^{-1}$.

V. COMPARISON WITH THE KERR EFFECT

All-optical nonlinearities such as bistability and saturable absorption have been previously considered as potential activation functions in ONNs [10], [31]. An alternative implementation of the activation function in Fig. 1(c) could consist of a nonlinear MZI, with one of its arms having a material with Kerr nonlinear optical response. The Kerr effect is a third-order optical nonlinearity which generates a change in the refractive index, and thus a nonlinear phase shift, which is proportional to the input pulse intensity. In this section we compare the electro-optic activation function introduced in the previous section [Fig. 1(c)] to such an alternative all-optical activation function using the Kerr effect, highlighting how the electro-optic activation can achieve a lower activation threshold.

Unlike the electro-optic activation function, the Kerr effect is lossless and has no latency because it arises from a nonlinear material response, rather than a feedforward circuit. A standard figure of merit for quantifying the strength of the Kerr effect in a waveguide is through the amount of nonlinear phase shift generated per unit input power per unit waveguide length. This is given mathematically by the expression

$$\Gamma_{\text{Kerr}} = \frac{2\pi}{\lambda_0} \frac{n_2}{A}, \qquad (12)$$

where n_2 is the nonlinear refractive index of the material and A is the effective mode area. Γ_{Kerr} ranges from 100 $(\text{W·m})^{-1}$ in chalcogenide to 350 $(\text{W·m})^{-1}$ in silicon [32]. An equivalent figure of merit for the electro-optic feedforward scheme can be mathematically defined as

$$\Gamma_{\text{EO}} = \pi \frac{\alpha \Re G}{V_\pi L}, \qquad (13)$$

where $V_\pi L$ is the phase modulator figure of merit. The figures of merit described in Eqs. (12)–(13) can be represented as an activation threshold (Eq. 8) via the relationship $P_{\text{th}} = \frac{\Delta\phi|_{\delta T=0.5}}{\Gamma L}$, for a given waveguide length, L where the electro-optic phase shift or nonlinear Kerr effect take place.

A comparison of Eq. 12 and Eq. 13 indicates that while the strength of the Kerr effect is largely fixed by waveguide design and material choice, the electro-optic scheme has several degrees
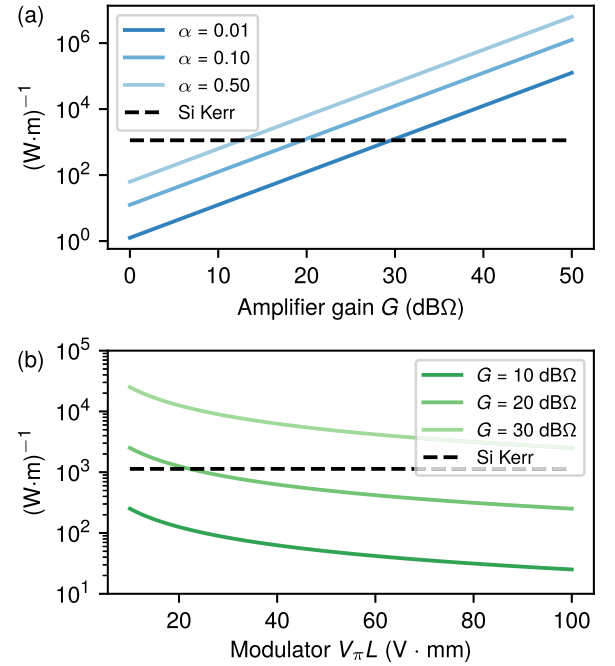

Fig. 4. Nonlinear parameter Γ_{EO} for the electro-optic activation as a function of (a) gain, G, for $\alpha = 0.50, 0.10$, and 0.01 and (b) modulator $V_\pi L$. The nonlinear parameter associated with the optical Kerr effect, Γ_{Kerr} in a Silicon waveguide of cross sectional area $A = 0.05\ \mu\text{m}^2$ corresponds to the black dotted line.

of freedom which allow it to potentially achieve a stronger nonlinear response. The first design parameter is the amount of power tapped off to the photodetector, which can be increased to generate a larger voltage at the phase modulator. However, increasing α also increases the *linear* signal loss through the activation which does not contribute to the nonlinear mapping between the input and output of the ONN. Therefore, α should be minimized as long as the optical power routed to the photodetector is large enough to be above the noise equivalent power level.

On the other hand, the product $\Re G$ determines the conversion efficiency of the detected optical power into an electrical voltage. Fig. 4(a) compares the nonlinearity strength of the electro-optic activation (blue lines) to that of an implementation using the Kerr effect in silicon (black dashed line) for several values of α, as a function of G. The responsivity is fixed at $\Re = 1.0$ A/W. We observe that tapping out 10% of the optical power requires a gain of 20 dBΩ to achieve a nonlinear phase shift equivalent threshold to that of a silicon waveguide where $A = 0.05\ \mu\text{m}^2$ for the same amount of input optical power. Tapping out only 1% of the optical power requires an additional 10 dBΩ of gain to maintain this equivalence. We note that the gain range considered in Fig. 4(a) is well within the regime of what has been demonstrated in integrated transimpedance amplifiers for optical receivers [24]–[26]. In fact, many of these systems have demonstrated much higher gain. In Fig. 4(a), the phase modulator $V_\pi L$ was fixed at 20 V·mm. However, because a lower $V_\pi L$ translates into an increased phase shift for a given applied voltage, this parameter can also be used to enhance the nonlinearity. Fig. 4(b) demonstrates the effect of changing the $V_\pi L$ for several values

of of G, again, with a fixed responsivity $\mathfrak{R} = 1.0$ A/W. This demonstrates that with a reasonable level of gain and phase modulator performance, the electro-optic activation function can trade off an increase in latency for a significantly lower optical activation threshold than the Kerr effect.

VI. MACHINE LEARNING TASKS

In this section, we apply the electro-optic activation function introduced above to several machine learning tasks. In Section VI-A, we simulate training an ONN to implement an exclusive-OR (XOR) logical operation. The network is modeled using `neuroptica` [33], a custom ONN simulator written in Python, which trains the simulated networks only from physically measurable field quantities using the on-chip backpropagation algorithm introduced in Ref. [23]. In Section VI-B, we consider the more complex task of using an ONN to classify handwritten digits from the Modified NIST (MNIST) dataset, which we model using the `neurophox` [34], [35] package and `tensorflow` [36], which computes gradients using automatic differentiation. In both cases, we model the values in the network as complex-valued quantities and represent the interferometer meshes as unitary matrices parameterized by phase shifters.

A. Exclusive-OR Logic Function

An exclusive-OR (XOR) is a logic function which takes two inputs and produces a single output. The output is *high* if only one of the two inputs is *high*, and *low* for all other possible input combinations. In this example, we consider a multi-input XOR which takes N input values, given by $x_1 \ldots x_N$, and produces a single output value, y. The input-output relationship of the multi-input XOR function is a generalization of the two-input XOR. For example, defining logical *high* and *low* values as 1 and 0, respectively, a four-input XOR has an output table indicated the desired values in Fig. 5(b). We select this task for the ONN to learn because it requires a non-trivial level of nonlinearity, meaning that it could not be implemented in an ONN consisting of only linear interferometer meshes.

The architecture of the ONN used to learn the XOR is shown schematically in Fig. 5(a). The network consists of L layers, with each layer constructed from an $N \times N$ unitary interferometer mesh followed by an array of N parallel electro-optic activation functions, with each element corresponding to the circuit in Fig. 1(c). After the final layer, the lower $N - 1$ outputs are dropped to produce a single output value which corresponds to y. Unlike the ideal XOR input-output relationship described above, for the XOR task learned by the ONN we normalize the input vectors such that they always have an L_2 norm of 1. This constraint is equivalent to enforcing a constant input power to the network. Additionally, because the activation function causes the optical power level to be attenuated at each layer, we take the *high* output state to be a value of 0.2, as shown in Fig. 1(b). The *low* output remains at a value of 0.0. An alternative to using a smaller amplitude for the output high state would be to add additional ports with fixed power biases to increase the total input power to the network, similarly to the XOR demonstrated in Ref. [23].
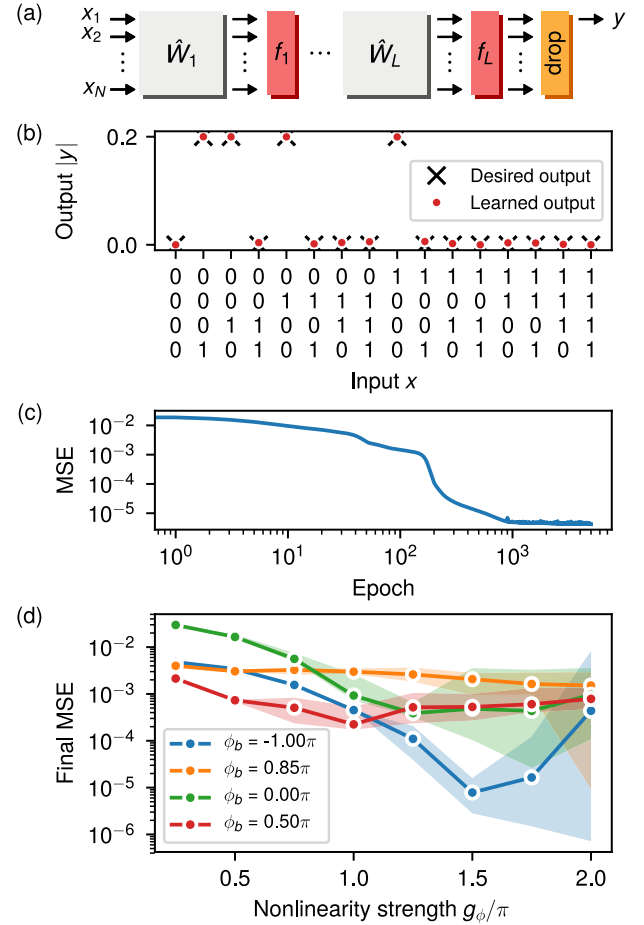

Fig. 5.   (a) Architecture of an L-layer ONN used to implement an N-input XOR logic function. (b) Red dots indicate the learned input-output relationship of the XOR for $N = 4$ on an 2-layer ONN. Electro-optic activation functions are configured with gain $g = 1.75\pi$ and biasing phase $\phi_b = \pi$. (c) Mean squared error (MSE) versus training epoch. (d) Final MSE after 5000 epochs averaged over 20 independent training runs vs activation function gain. Different lines correspond to the responses shown in Fig. 2, with $\phi_b = 1.00\pi, 0.85\pi, 0.00\pi$, and 0.50π. Shaded regions correspond to the range (minimum and maximum) final MSE from the 20 training runs.

In Fig. 5(b) we show the four-input XOR input-output relationship which was learned by a two-layer ONN. The electro-optic activation functions were configured to have a gain of $g = 1.75\pi$ and biasing phase of $\phi_b = \pi$. This biasing phase configuration corresponds to the ReLU-like response shown in Fig. 2(a). The black markers indicate the desired output values while the red circles indicate the output learned by the two-layer ONN. Fig. 5(b) indicates excellent agreement between the learned output and the desired output. The evolution of the mean squared error (MSE) between the ONN output and the desired output during training confirms this agreement, as shown in Fig. 5(c), with a final MSE below 10^{-5}.

To train the ONN, a total of $2^N = 16$ training examples were used, corresponding to all possible binary input combinations along the x-axis of Fig. 5(b). All 16 training examples were fed through the network in a batch to calculate the mean squared error (MSE) loss function. The gradient of the loss function with respect to each phase shifter was computed by backpropagating

the error signal through the network to calculate the loss sensitivity at each phase shifter [23]. The above steps were repeated until the MSE converged, as shown in Fig. 5(c). Only the phase shifter parameters were optimized by the training algorithm, while all parameters of the activation function were unchanged.

To demonstrate that the nonlinearity provided by the electro-optic activation function is essential for the ONN to successfully learn the XOR, in Fig. 5(d) we plot the final MSE after 5000 training epochs, averaged over 20 independent training runs, as a function of the activation function gain, g_ϕ. The shaded regions indicates the minimum and maximum range of the final MSE over the 20 training runs. The four lines shown in Fig. 5(d) correspond to the four activation function bias configurations shown in Fig. 2.

For the blue curve in Fig. 5(d), which corresponds to the ReLU-like activation, we observe a clear improvement in the final MSE with an increase in the nonlinearity strength. We also observe that for very high nonlinearity, above $g_\phi = 1.5\pi$, the range between the minimum and maximum final MSE broadens and the mean final MSE increases. However, the best case (minimum) final MSE continues to decrease, as indicated by the lower border of the shaded blue region. This trend indicates that although increasing nonlinearity improves the ONN's ability to learn the XOR function, very high levels of nonlinearity may also prevent the training algorithm from converging.

A trend of decreasing MSE with increasing nonlinearity is also observed for the activation corresponding to the green curve in Fig. 5(d). However, the range of MSE values begins to broaden at a lower value of $g_\phi = 1.0\pi$. Such broadening may be a result of the changing slope in the activation function output, as shown in Fig. 2(e). For the activation functions corresponding to the red and orange curves in Fig. 5(d), the final MSE decreases somewhat with an increase in g_ϕ, but generally remains much higher than the other two activation function responses. We conclude that these two responses are not as well suited for learning the XOR function. Overall, these results demonstrate that the flexibility of our architecture to achieve specific forms of nonlinear activation functions is important for the successful operation of an ONN.

B. Handwritten Digit Classification

The second task we consider for demonstrating the activation function is classifying images of handwritten digits from the MNIST dataset, which has become a standard benchmark problem for ANNs [37]. The dataset consists of 70,000 grayscale 28×28 pixel images of handwritten digits between 0 and 9. Several representative images from the dataset are shown in Fig. 6(a).

To reduce the number of input parameters, and hence the size of the neural network, we use a preprocessing step to convert the images into a Fourier-space representation. Specifically, we compute the 2D Fourier transform of the images which is defined mathematically as $c(k_x, k_y) = \sum_{m,n} e^{jk_x m + jk_y n} g(m, n)$, where $g(m, n)$ is the gray scale value of the pixel at location (m, n) within the image. The

amplitudes of the Fourier coefficients $c(k_x, k_y)$ are shown below their corresponding images in Fig. 6(a). These coefficients are generally complex-valued, but because the real-space map $g(m, n)$ is real-valued, the condition $c(k_x, k_y) = c^*(-k_x, -k_y)$ applies.

We observe that the Fourier-space profiles are mostly concentrated around small k_x and k_y, corresponding to the center region of the profiles in Fig. 6(a). This is due to the slowly varying spatial features in the images. We can therefore expect that most of the information is carried by the small-k Fourier components, and with the goal of decreasing the input size, we can restrict the data to N coefficients with the smallest $k = \sqrt{k_x^2 + k_y^2}$. An additional advantage of this preprocessing step is that it reduces the computational resources required to perform the training process because the neural network dimension does not need to accommodate all $28^2 = 784$ pixel values as inputs.

Fourier preprocessing is particularly relevant for ONNs for two reasons. First, the Fourier transform has a straightforward implementation in the optical domain using techniques from Fourier optics involving standard components such as lens and spatial filters [38]. Second, this approach allows us to take advantage of the fact that ONNs are *complex*-valued functions. That is to say, the N complex-valued coefficients $c(k_x, k_y)$ can be handled by an N-dimensional ONN, whereas to handle the same input using a real-valued neural network requires a twice larger dimension. The ONN architecture used in our demonstration is shown schematically in Fig. 6(a). The N Fourier coefficients closest to $k_x = k_y = 0$ are fed into an optical neural network consisting of L layers, after which a drop-mask reduces the final output to 10 components. The intensity of the 10 outputs are recorded and normalized by their sum, which creates a probability distribution that may be compared with the one-hot encoding of the digits from 0 to 9. The loss function is defined as the cross-entropy between the normalized output intensities and the correct one-hot vector.

During each training epoch, a subset of 60,000 images from the dataset were fed through the network in batches of 500. The remaining 10,000 image-label pairs were used to form a test dataset. For a two-layer network with $N = 16$ Fourier components, Fig. 6(b) compares the classification accuracy over the training dataset (solid lines) and testing dataset (dashed lines) while Fig. 6(b) compares the cross entropy loss during optimization. The blue curves correspond to an ONN with no activation function (e.g. a linear optical classifier) and the orange curves correspond to an ONN with the electro-optic activation function configured with $g_\phi = 0.05\pi$, $\phi_b = 1.00\pi$, and $\alpha = 0.1$. The gain setting in particular was selected heuristically. We observe that the nonlinear activation function results in a significant improvement to the ONN performance during and after training. The final validation accuracy for the ONN with the activation function is 93%, which amounts to an 8% difference as compared to the linear ONN which achieved an accuracy of 85%.

The confusion matrix computed over the testing dataset is shown in Fig. 6(d). We note that the prediction accuracy of 93% is high considering that only $N = 16$ complex Fourier components were used, and the network is parameterized by only
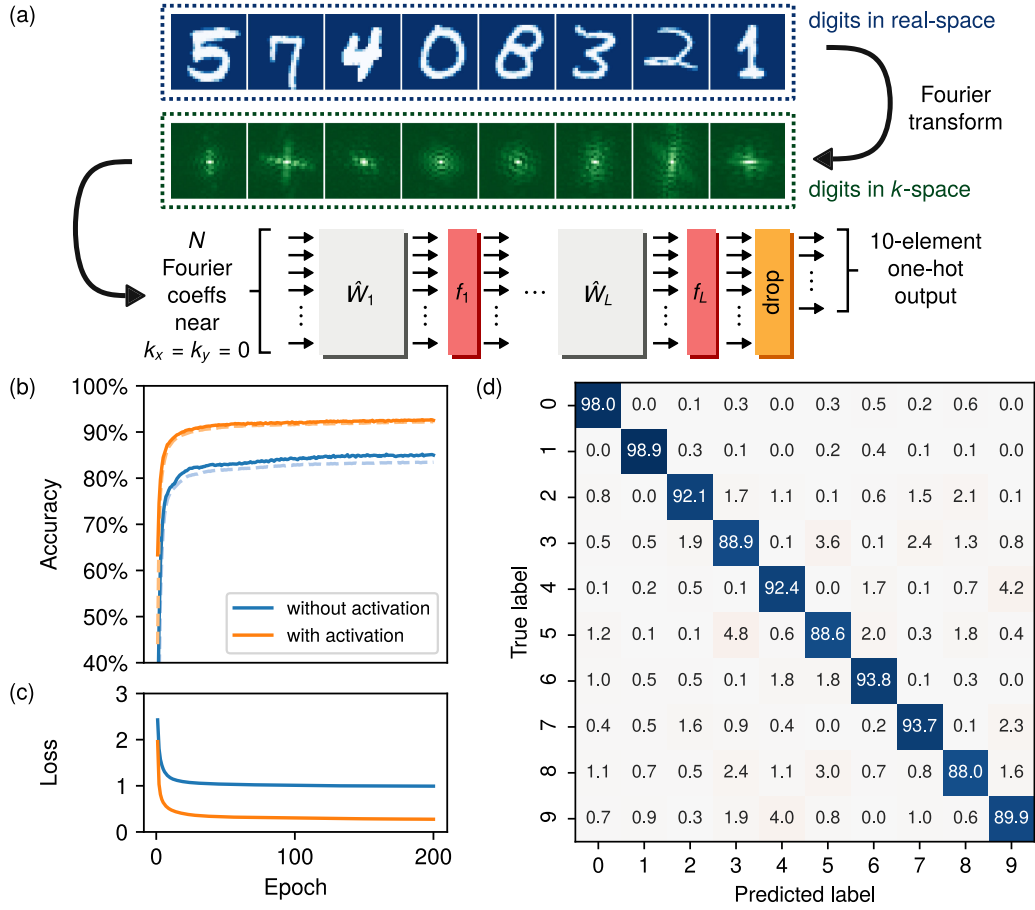
Fig. 6.   (a) Schematic of an optical image recognition setup based on an ONN. Images of handwritten numbers from the MNIST database are preprocessed by converting from real-space to k-space and selecting N Fourier coefficients associated with the smallest magnitude k-vectors. (b) Test accuracy (solid lines) and training accuracy (dashed lines) during training for a two layer ONN without activation functions (blue) and with activation functions (orange). $N = 16$ Fourier components were used as inputs to the ONN and each vector was normalized such that its L_2 norm is unity. The activation function parameters were $g_\phi = 0.05\pi$ and $\phi_b = 1.00\pi$. (c) Cross entropy loss during training. (d) Confusion matrix, specified in percentage, for the trained ONN with the electro-optic activation function.

TABLE III
ACCURACY ON THE MNIST TESTING DATASET AFTER OPTIMIZATION

# Layers	Without activation	With activation	
		Untrained	Trained*
1	85.00%	89.80%	89.38%
2	85.83%	92.98%	92.60%
3	85.16%	92.62%	93.89%

*The phase gain, g_ϕ, of each layer was optimized during training.

$2 \times N^2 \times L = 1024$ free parameters. Moreover, this prediction accuracy is comparable with the 92.6% accuracy achieved in a fully-connected linear classifier with 4010 free parameters taking *all* of the $28^2 = 784$ real-space pixel values as inputs [37]. Finally, in Table III we show that the accuracy can be further improved by including a third layer in the ONN and by making the activation function gain a trainable parameter. This brings the testing accuracy to 94%. Based on the parameters from Table I and the scaling from Table II, the 3 layer handwritten digit classification system would consume 4.8 W while performing 7.7×10^{12} MAC/sec. Its prediction latency would be 1.5 ns.

VII. CONCLUSION

In conclusion, we have introduced an architecture for synthesizing optical-to-optical nonlinearities and demonstrated its use as a nonlinear activation function in a feed forward ONN. Using numerical simulations, we have shown that such activation functions enable an ONN to be successfully applied to two machine learning benchmark problems: (1) learning a multi-input XOR logic function, and (2) classifying handwritten numbers from the MNIST dataset. Rather than using all-optical nonlinearities, our activation architecture uses intermediate signal pathways in the electrical domain which are accessed via photodetectors and phase modulators. Specifically, a small portion of the optical input power is tapped out which undergoes analog processing before modulating the remaining portion of the same optical signal. Whereas all-optical nonlinearities have largely fixed responses, a benefit of the electro-optic approach demonstrated here is that signal amplification in the electronic domain can overcome the need for high optical signal powers to achieve a significantly lower activation threshold. For example, we show that a phase modulator V_π of 10 V and an optical-to-electrical

conversion gain of 57 dBΩ, both of which are experimentally feasible, result in an optical activation threshold of 0.1 mW. We note that this nonlinearity is compatible with the *in situ* training protocol proposed in Ref. [23], which is applicable to arbitrary activation functions.

Our activation function architecture can utilize the same integrated photodetector and modulator technologies as the input and output layers of a fully-integrated ONN. This means that an ONN using this activation suffers no reduction in processing speed, despite using analog electrical components. The only trade off made by our design is an increase in latency due to the electro-optic conversion process. However, we find that an ONN with dimension $N = 100$ has a total prediction latency of 2.4 ns/layer, with approximately equal contributions from the propagation of optical pulses through the interferometer mesh and from the electro-optic activation function. Conservatively, we estimate the energy consumption of an ONN with this activation function to be 100 fJ/MAC, but this figure of merit could potentially be reduced by orders of magnitude using highly efficient modulators and amplifier-free optoelectronics [29].

Finally, we emphasize that in our activation function, the majority of the signal power remains in the optical domain. There is no need to have a new optical source at each nonlinear layer of the network, as is required in previously demonstrated electro-optic neuromorphic hardware [14], [21], [39] and reservoir computing architectures [40], [41]. Additionally, each activation function in our proposed scheme is a standalone analog circuit and therefore can be applied in parallel. While we have focused here on the application of our architecture as an activation function in a feedforward ONN, the synthesis of low-threshold optical nonlinearlities using this circuit could be of broader interest for optical computing as well as microwave photonic signal processing applications.

References

[1] V. K. Pallipuram, M. Bhuiyan, and M. C. Smith, "A comparative study of GPU programming models and architectures using neural networks," *J. Supercomput.*, vol. 61, no. 3, pp. 673–718, Sep. 2012.

[2] J. M. Shainline, S. M. Buckley, R. P. Mirin, and S. W. Nam, "Superconducting optoelectronic circuits for neuromorphic computing," *Phys. Rev. Appl.*, vol. 7, no. 3, Mar. 2017, Art. no. 034013.

[3] B. J. Shastri *et al.*, "Principles of neuromorphic photonics," *Encyclopedia of Complexity and Systems Science.* Berlin, Germany: Springer 2018, pp. 1–37.

[4] F. D. Coarer *et al.*, "All-optical reservoir computing on a photonic chip using silicon-based ring resonators," *IEEE J. Sel. Topics Quantum Electron.*, vol. 24, no. 6, Nov. 2018, Art. no. 7600108.

[5] 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.*, vol. 8, no. 1, Aug. 2018, Art. no. 12324.

[6] 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.

[7] J. Capmany and D. Novak, "Microwave photonics combines two worlds," *Nature Photon.*, vol. 1, no. 6, pp. 319–330, 2007.

[8] D. Marpaung *et al.*, "Integrated microwave photonics," *Laser Photon. Rev.*, vol. 7, no. 4, pp. 506–538, Jul. 2013.

[9] P. Ghelfi *et al.*, "A fully photonics-based coherent radar system," *Nature*, vol. 507, no. 7492, pp. 341–345, Mar. 2014.

[10] Y. S. Abu-Mostafa and D. Psaltis, "Optical neural computers," *Sci. Am.*, vol. 256, no. 3, pp. 88–95, Mar. 1987.

[11] D. Psaltis, D. Brady, X.-G. Gu, and S. Lin, "Holography in artificial neural networks," *Nature*, vol. 343, no. 6256, pp. 325–330, Jan. 1990.

[12] Y. Shen *et al.*, "Deep learning with coherent nanophotonic circuits," *Nature Photon.*, vol. 11, no. 6, pp. 441–447, Jun. 2017.

[13] D. A. B. Miller, "Self-configuring universal linear optical component," *Photon. Res.*, vol. 1, no. 1, pp. 1–15, Jun. 2013.

[14] A. N. Tait *et al.*, "Neuromorphic photonic networks using silicon photonic weight banks," *Sci. Rep.*, vol. 7, no. 1, Aug. 2017, Art. no. 7430.

[15] M. Radulaski, R. Bose, T. Tran, T. Van Vaerenbergh, D. Kielpinski, and R. G. Beausoleil, "Thermally tunable hybrid photonic architecture for nonlinear optical circuits," *ACS Photon.*, vol. 5, no. 11, pp. 4323–4329, Nov. 2018.

[16] Q. Bao, H. Zhang *et al.*, "Monolayer graphene as a saturable absorber in a mode-locked laser," *Nano Res.*, vol. 4, no. 3, pp. 297–307, Mar. 2011.

[17] N. H. Park *et al.*, "Monolayer graphene saturable absorbers with strongly enhanced evanescent-field interaction for ultrafast fiber laser mode-locking," *Opt. Exp.*, vol. 23, no. 15, pp. 19806–19812, Jul. 2015.

[18] X. Jiang *et al.*, "Low-dimensional nanomaterial saturable absorbers for ultrashort-pulsed waveguide lasers," *Opt. Mater. Express*, vol. 8, no. 10, pp. 3055–3071, Oct. 2018.

[19] A. L. Lentine and D. A. B. Miller, "Evolution of the SEED technology: Bistable logic gates to optoelectronic smart pixels," *IEEE J. Quantum Electron.*, vol. 29, no. 2, pp. 655–669, Feb. 1993.

[20] A. Majumdar and A. Rundquist, "Cavity-enabled self-electro-optic bistability in silicon photonics," *Opt. Lett.*, vol. 39, no. 13, pp. 3864–3867, Jul. 2014.

[21] A. N. Tait *et al.*, "Silicon Photonic modulator neuron," *Phys. Rev. Appl.*, vol. 11, no. 6, Jun. 2019, Art. no. 064043.

[22] E. Trentin, "Networks with trainable amplitude of activation functions," *Neural Netw.*, vol. 14, no. 4, pp. 471–493, May 2001.

[23] T. W. Hughes, M. Minkov, Y. Shi, and S. Fan, "Training of photonic neural networks through in situ backpropagation and gradient measurement," *Optica*, vol. 5, no. 7, pp. 864–871, Jul. 2018.

[24] M. N. Ahmed, J. Chong, and D. S. Ha, "A 100 Gb/s transimpedance amplifier in 65 nm CMOS technology for optical communications," in *Proc. IEEE Int. Symp. Circuits Syst.*, Jun. 2014, pp. 1885–1888.

[25] K. T. Settaluri, C. Lalau-Keraly, E. Yablonovitch, and V. Stojanović, "First principles optimization of opto-electronic communication links," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 64, no. 5, pp. 1270–1283, May 2017.

[26] K. Nozaki, S. Matsuo, A. Shinya, and M. Notomi, "Amplifier-free bias-free receiver based on low-capacitance nanophotodetector," *IEEE J. Sel. Topics Quantum Electron.*, vol. 24, no. 2, Mar. 2018, Art. no. 4900111.

[27] S. K. Selvaraja, *et al.*, "Highly uniform and low-loss passive silicon photonics devices using a 300mm CMOS platform," presented at the Opt. Fiber Commun. Conf., Mar. 2014, Paper Th2A.33.

[28] J. Cardenas *et al.*, "Wide-bandwidth continuously tunable optical delay line using silicon microring resonators," *Opt. Express*, vol. 18, no. 25, pp. 26525–26534, Dec. 2010.

[29] K. Nozaki *et al.*, "Femtofarad optoelectronic integration demonstrating energy-saving signal conversion and nonlinear functions," *Nature Photon.*, vol. 13, pp. 454–459, Apr. 2019.

[30] G. Yu *et al.*, "A low-noise high-gain transimpedance amplifier with high dynamic range in 0.13 μm CMOS," in *Proc. IEEE Int. Symp. Radio-Freq. Integr. Technol.*, Nov. 2012, pp. 37–40.

[31] Y. Shen *et al.*, "Supplementary information: Deep learning with coherent nanophotonic circuits," *Nature Photon.*, vol. 11, no. 7, pp. 441–446, Jun. 2017.

[32] C. Koos, L. Jacome, C. Poulton, J. Leuthold, and W. Freude, "Nonlinear silicon-on-insulator waveguides for all-optical signal processing," *Opt. Express*, vol. 15, no. 10, pp. 5976–5990, May 2007.

[33] "Neuroptica: An optical neural network simulator," [Online]. Available: https://github.com/fancompute/neuroptica

[34] "Neurophox: A simulation framework for unitary neural networks and photonic devices," [Online]. Available: https://github.com/solgaardlab/neurophox

[35] S. Pai, B. Bartlett, O. Solgaard, and D. A. B. Miller, "Matrix optimization on universal unitary photonic devices," *Phys. Rev. Appl.*, vol. 11, no. 6, Jun. 2019, Art. no. 064044.

[36] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. [Online]. Available: tensorflow.org

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

[38] J. W. Goodman, *Introduction to Fourier Optics*. Englewood, CO, USA: Roberts and Company Publishers, 2005.

[39] H. Peng, M. A. Nahmias, T. F. de Lima, A. N. Tait, and B. J. Shastri, "Neuromorphic photonic integrated circuits," *IEEE J. Sel. Topics Quantum Electron.*, vol. 24, no. 6, Nov. 2018, Art. no. 6101715.

[40] L. Larger *et al.*, "Photonic information processing beyond Turing: An optoelectronic implementation of reservoir computing," *Opt. Express*, vol. 20, no. 3, pp. 3241–3249, Jan. 2012.

[41] F. Duport, A. Smerieri, A. Akrout, M. Haelterman, and S. Massar, "Fully analogue photonic reservoir computer," *Sci. Rep.*, vol. 6, Mar. 2016, Art. no. 22381.

Ian A. D. Williamson (M'10) received the B.S., M.S., and Ph.D. degrees in electrical engineering from The University of Texas at Austin, Austin, TX, USA, in 2010, 2014, and 2017, respectively. He is currently a Postdoctoral Research Fellow with the Department of Electrical Engineering and the Edward L. Ginzton Laboratory, Stanford University, Stanford, CA, USA.

His research interests include signal processing, machine learning, and computing with optical and other analog hardware platforms.

Tyler W. Hughes received the B.S. degree in physics from the University of Michigan, Ann Arbor, MI, USA, in 2013, and the M.S. degree in 2016 from the Department of Applied Physics, Stanford University, Stanford, CA, USA. He is currently working toward the Ph.D. degree with Stanford University.

After receiving the B.S. degree, he became a Research Assistant with the National University of Singapore's Centre for Quantum Technologies. His research interests include electromagnetic inverse design, photonic hardware platforms for machine learning, and laser-driven particle accelerators on a chip.

Momchil Minkov received the B.S. degree in physics from Jacobs University Bremen, Bremen, Germany, in 2010, and the M.S. and Ph.D. degrees in physics from Swiss National Polytechnic Institute, Lausanne, Switzerland, in 2012 and 2016, respectively.

He is currently a Postdoctoral Research Associate with Stanford University, Stanford, CA, USA, under a fellowship from the Swiss National Science Foundation. He has coauthored more than 30 peer-reviewed articles. His research interests include photonic crystals and their applications, topological photonics, and machine learning with optics and photonics.

Ben Bartlett received the B.S. degree in physics and computer science from the California Institute of Technology, Pasadena, CA, USA, in 2017. He is currently working toward the Ph.D. degree in applied physics at Stanford University, Stanford, CA, USA.

His research interests include optical and quantum information processing, nanophotonics, machine learning, and quantum optics.

Sunil Pai received the B.S. degree in physics and the M.S. degree in computer science from Stanford University, Stanford, CA, USA, in 2015 and 2016, respectively. He is currently working toward the Ph.D. degree in electrical engineering with Stanford University, Stanford, CA, USA.

His research interests include machine learning, photonics, and quantum optics.

Shanhui Fan (F'11) received the Ph.D. degree in theoretical condensed matter physics from the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in 1997.

He is currently a Professor in electrical engineering, a Professor in applied physics (by courtesy), a Senior Fellow with the Precourt Institute of Energy, and the Director of the Edward L. Ginzton Laboratory, Stanford University, Stanford, CA, USA. He was a Research Scientist with the Research Laboratory of Electronics, MIT. He has authored or coauthored more than 400 refereed journal articles that were cited more than 35 000 times according to the Web of Sciences. He has given more than 300 invited talks and holds 62 U.S. patents. His research interests include fundamental studies of solid-state and photonic structures and devices, especially photonic crystals, plasmonics, and metamaterials, and their applications in information and energy technology.

Dr. Fan is a Fellow of the American Physical Society, the Optical Society of America, and the SPIE. He was the recipient of the National Science Foundation Career Award in 2002, the David and Lucile Packard Fellowship in Science and Engineering in 2003, the National Academy of Sciences Award for Initiative in Research in 2007, the Adolph Lomb Medal from the Optical Society of America in 2007, and the Vannevar Bush Faculty Fellowship in 2017. He has been listed as a highly cited researcher in physics by Thomson Reuters since 2015.