Digital Electronics and Analog Photonics for Convolutional Neural Networks (DEAP-CNNs)

Viraj Bangari,* Bicky A. Marquez, Heidi B. Miller, and Bhavin J. Shastri†
Department of Physics, Engineering Physics & Astronomy, Queen's University, Kingston, ON KL7 3N6, Canada

Alexander N. Tait
National Institute of Standards and Technology (NIST), Boulder, Colorado 80305, USA

Mitchell A. Nahmias, Thomas Ferreira de Lima, Hsuan-Tung Peng, and Paul R. Prucnal
Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA

(Dated: July 3, 2019)

Convolutional Neural Networks (CNNs) are powerful and highly ubiquitous tools for extracting features from large datasets for applications such as computer vision and natural language processing. However, a convolution is a computationally expensive operation in digital electronics. In contrast, neuromorphic photonic systems, which have experienced a recent surge of interest over the last few years, propose higher bandwidth and energy efficiencies for neural network training and inference. Neuromorphic photonics exploits the advantages of optical electronics, including the ease of analog processing, and busing multiple signals on a single waveguide at the speed of light. Here, we propose a Digital Electronic and Analog Photonic (DEAP) CNN hardware architecture that has potential to be 2.8 to 14 times faster while maintaining the same power usage of current state-of-the-art GPUs.

I. INTRODUCTION

The success of CNNs for large-scale image recognition has stimulated research in developing faster and more accurate algorithms for their use. However, CNNs are computationally intensive and therefore results in long processing latency. One of the primary bottlenecks is computing the matrix multiplication required for forward propagation. In fact, over 80% of the total processing time is spent on the convolution [1]. Therefore, techniques that improve the efficiency of even forward-only propagation are in high demand and researched extensively [2, 3].

In this work, we present a complete digital electronic and analog photonic (DEAP) architecture capable of performing highly efficient CNNs for image recognition. The competitive MNIST handwriting dataset[4] is used as a benchmark test for our DEAP CNN. At first, we train a standard two-layer CNN offline, after which network parameters are uploaded to the DEAP CNN. Our scope is limited to the forward propagation, but includes power and speed analyses of our proposed architecture.

Due to their speed and energy efficiency, photonic neural networks have been widely investigated from different approaches that can be grouped into three categories: (1) reservoir computing [5–8]; reconfigurable architectures based on (2) ring-resonators [9–12], and (3) Mach-Zehnder interferometers [13, 14]. Reservoir computing in the discrete photonic domain successfully implement neural networks for fast information processing, however the predefined random weights of their hidden layers cannot be modified [8].

An alternative approach uses silicon photonics to design fully programmable neural networks [15], using a so-called broadcast-and-weight protocol [10–12]. This protocol is capable of implementing reconfigurable, recurrent and feedforward neural network models, using a bank of tunable silicon microring resonators (MRRs) that recreate on-chip synaptic weights. Therefore, such a protocol allows it to emulate physical neurons. Mach-Zehnder interferometers have been also used to model synaptic-like connections of physical neurons [14]. The advantage of the former approach over the latter is that it has already demonstrated fan-in, inhibition, time-resolved processing, and autaptic cascadability [12]. The DEAP CNN design is therefore compatible with mainstream silicon photonic device platforms. This approach leverages the advances in silicon photonics that have recently progressed to the level of sophistication required for large-scale integration. Furthermore, this proposed architecture allows the implementation of multi-layer networks to implement the deep learning framework.

Inspired by the work of Mehrabian et al. [16], which lays out a potential architecture for photonic CNNs with DRAM, buffers, and microring resonators, our design goes a step further by considering specific input representation, as well as an example of how an algorithm for tasks such as MNIST handwritten digit recognition can be mapped to photonics. Moreover, we consider summation of multi-channel inputs, multi-dimensional kernels, the limitations on weights being between 0 and 1, and the architecture for the depth of kernel or inputs.

This work is divided in five sections: Following this introduction, in section (II), we describe convolutions as used in the field of signal processing. Then, we intro-

* viraj.bangari@queensu.ca
† shastri@ieee.org

duce silicon photonic devices to perform convolutions in photonics. Section (III) introduces a hardware inspired algorithm to perform such full photonic convolutions. In Section (IV), we utilize our previously described architecture to build a two-layers DEAP CNN for MNIST handwritten digit recognition. Finally, in section (V), we show an energy-speed benchmark test, where we compare the performance of DEAP with the empirical dataset DeepBench [17]. Note, we have made the high level simulator and mapping tool for the DEAP architecture publicly available [18].

II. CONVOLUTIONS AND PHOTONICS

II.1. Convolutions Background

A convolution of two discrete domain functions f and g is defined by:

$$(f * g)[t] = \sum_{t=-\infty}^{\infty} f[\tau]g[t-\tau], \tag{1}$$

where $(f*g)$ represents a weighted average of the function $f[\tau]$ when it is weighting by $g[-\tau]$ shifted by t. The weighting function $g[-\tau]$ emphasizes different parts of the input function $f[\tau]$ as t changes.

In digital image processing, a similar process is followed. The convolution of an image A with a kernel F produces a convolved image O. An image is represented as a matrix of numbers with dimensionality $H \times W$, where H and W are the height and width of the image, respectively. Each element of a matrix represents the intensity of a pixel at that particular spatial location. A kernel is a matrix of real numbers with dimensionality $R \times R$. The value of a particular convolved pixel is defined by:

$$O_{i,j} = \sum_{k=1}^{R}\sum_{l=1}^{R} F_{k,l}A_{i+k,j+l}. \tag{2}$$

Using matrix slicing notation, Eq. (2) can be represented as a dot product of two vectorized matrices:

$$O_{i,j} = \mathrm{vec}(F)^T \cdot \mathrm{vec}((A_{m,n})_{n\in[j,j+R]}^{m\in[i,i+R]})^T. \tag{3}$$

A convolution reduces the dimensionality of the input image to $(H-R+1)\times(W-R+1)$, so a padding of zero values is normally applied around the edges of the input image to counteract this. A schematic illustration of a convolution in digital image processing is shown at the top of Fig. 1.

When convolutions are used to perform parallel matrix multiplications in neural networks such as CNNs, a convolution operation is defined as:

$$O_{i,j} = \mathrm{vec}(F)^T \cdot \mathrm{vec}((A_{m,n,k})_{\substack{n\in[jS,jS+R]\\k\in[1,D]}}^{m\in[iS,iS+R]})^T, \tag{4}$$

where the input A has dimensionality $H\times W\times D$, kernel F has dimensionality $R\times R\times D$ and D refers to the number of channels within the input image. The additional

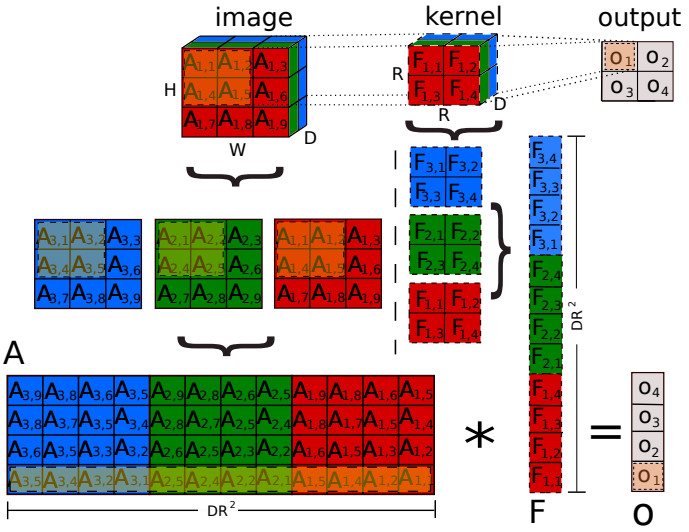

Figure 1. Schematic illustration of a convolution. At the top of the figure, an input image is represented as a matrix of numbers with dimensionality $H \times W \times D$ where H, W and D are the height, width and depth of the image, respectively. Each element $A_{i,j}$ of A represents the intensity of a pixel at that particular spatial location. The kernel F is a matrix with dimensionality $R \times R \times D$, where each element $F_{i,j}$ is defined as a real number. The kernel is slid over the image by using a stride S equal to one. As the image has multiple channels (or depth) D, the same kernel is applied to each channel. Assuming $H = W$, the overall output dimensionality is $(H-R+1)^2$. The bottom of the figure shows how a convolution operation generalized into a single matrix-matrix multiplication. where the kernel F is transformed into a vector $\mathbf{F}$ with DR^2 elements, and the image A is transformed into a matrix $\mathbf{A}$ of dimensionality $DR^2 \times (H-R+1)^2$. Therefore, the output is represented by a vector with $(H-R+1)$ elements.

Table I. Summary of Convolutional Parameters

Parameter	Meaning
N	Number of input images
H	Height of input image including padding
W	Width of input image including padding
D	Number of input channels
R	Edge length of kernel
K	Number of kernels
S	Stride

parameter S is referred to as the "stride" of the convolution. This convolution is similar to Eq. (3), except that the outputs from each channel are summed together in the end, and that the stride parameter is always equal to 1 in image processing. The dimensionality of the output feature is:

$$\left\lceil \frac{H-R}{S}+1 \right\rceil \times \left\lceil \frac{W-R}{S}+1 \right\rceil \times K, \tag{5}$$

where K is the number of different kernels applied to an image, and $\lceil\cdot\rceil$ is the ceiling function. Table (I) contains

a summary of all the convolutional parameters described so far.

One of the challenges with convolutions is that they are computationally intensive operations, taking up 86% to 94% of execution time for CNNs [1]. For heavy workloads, convolutions are typically run on graphical processing units (GPUs), as they are able to perform many mathematical operations in parallel. A GPU is a specialized hardware unit that is capable of performing a single mathematical operation on large amounts of data at once. This parallelization allow GPUs to compute matrix-matrix multiplication at speeds much higher than a CPU [19]. The convolution operation can be generalized into a single matrix-matrix multiplication [20]. This is shown at the bottom of Fig. 1, where the kernel F is transformed into a vector $\mathbf{F}$ with dimensionality $KDR^2 \times 1$, and the image is transformed into a matrix $\mathbf{A}$ of dimensionality $KDR^2 \times \left\lceil \frac{H-R}{S}+1 \right\rceil \left\lceil \frac{W-R}{S}+1 \right\rceil K$. Therefore, the output is represented by a vector with $\left\lceil \frac{H-R}{S}+1 \right\rceil \left\lceil \frac{W-R}{S}+1 \right\rceil K$ elements; where in this particular case $K=1$, $S=1$ and $H=W$.

II.2. Silicon Photonics Background

An emerging alternative to GPU computing is optical computing using silicon photonics for ultrafast information processing. Silicon photonics is a technology that allows for the implementation of photonic circuits by using the existing complementary-metal-oxide-semiconductor (CMOS) platform for electronics [21]. In recent years, the silicon photonic based broadcast-and-weight architecture has been shown to perform multiply-accumulate operations at frequencies up to five times faster than conventional electronics [22]. Therefore, there is motivation to explore how photonics can be used to perform convolutions, and how it compares to GPU-based implementations.

MRRs are the essential devices of our approach. A MRR is a circular waveguide that is coupled with either one or two waveguides. Such silicon waveguides can be manufactured to have a width of 500 nm while having a thickness of 220 nm. These waveguides have a bend radius of 5 μm and can support TE and TM polarized wavelengths between 1.5 μm and 1.6 μm [21]. The single waveguide configuration is called an all-pass MRR, see Fig. 2(a).

The light from the waveguide is transferred into the ring via a directional coupler and then recombined. The effective index of refraction between the waveguide and the MRR and the circumference of the MRR cause the recombined wave to have a phase shift, thereby interfering with the intensity of original light. The transfer function of the intensity of the light coming out through port with the light going into the input port of the all-pass resonator is described by:

$$T_n(\phi) = \frac{a^2 - 2ra\cos(\phi) + r^2}{1 - 2ra\cos(\phi) + (ar)^2}. \tag{6}$$

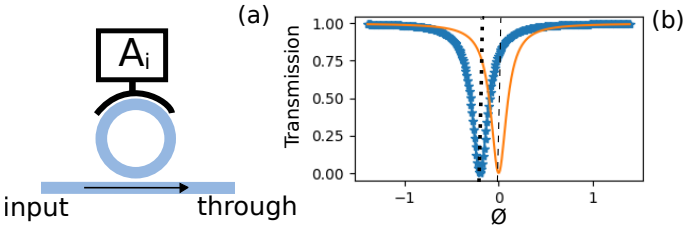

Figure 2. (a) All-pass MRR and (b) transfer function: the orange curve represents the Lorentzian line shape described by Eq. (6), centered in the initial phase where MRR is in resonance with the incoming light. The blue triangle curve shows how such phase can be modified by heating the MRR via the application of a current through A_i.

The parameter r is the self-coupling coefficient, and a defines the propagation loss from the ring and the directional coupler. The phase ϕ depends on the wavelength λ of the light and radius d of the MRR [23]:

$$\phi = \frac{4\pi^2 d n_{eff}}{\lambda}, \tag{7}$$

where n_{eff} is the effective index of refraction between the ring and waveguide. The value of n_{eff} can be modified to indirectly change the resonance peak. Such tuning is usually made by applying current to the ring proportional to the variable A_i. This process heats the ring, yielding a shift of the resonance peak. Figure 2(b) shows an example of such tuning: the orange curve represents the Lorentzian line shape described by Eq. (6), centered in the initial phase of the ring resonator, indicating that the MRR is in resonance with the incoming light. The blue triangle curve shows how such phase can be modified by heating the MRR.

The phase for an all-pass resonator corresponding to a particular intensity modulation value can be computed by using Eq. (6):

$$\phi_i = \arccos\left[\frac{A_i(1+(ar)^2) - a^2 - r^2}{2ra(1-A_{i,j})}\right], \tag{8}$$

resulting in a modulated intensity equal to A_i:

$$I_{mod} = T_n(\phi_i)|E_0|^2 = A_i, \tag{9}$$

where E_0 is amplitude of the electric field.

An alternative double waveguide configuration is called the add-drop MRR. The transfer function of the through port light intensity with respect to the input light is:

$$T_p(\phi) = \frac{(ar)^2 - 2r^2\cos(\phi) + r^2}{1 - 2r^2\cos(\phi) + (r^2a)^2}; \tag{10}$$

and the transfer function of the drop port light intensity with respect to the input light is:

$$T_d(\phi) = \frac{(1-r)^2 a}{1 - 2r^2\cos(\phi) + (r^2a)^2}. \tag{11}$$

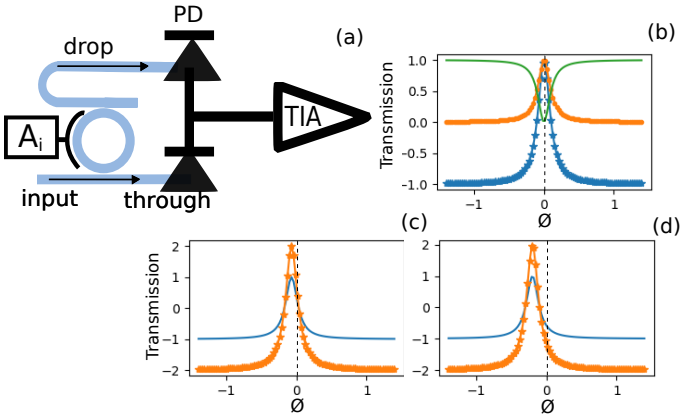

Figure 3. (a) Add-drop configuration and O/E conversion and amplification. (b) Output of the balanced photodiode, the transfer function of $T_p - T_d$. Orange circle and green curves are drop and through ports, described by Eqs. (11) and (10), respectively. In panels (c) and (d), the phase shifted ($\phi + 0.2$) blue curves show how such positive and negative kernel values from the drop and the through outputs, respectively. The orange triangle curves show how those values can be amplified by a factor of two using a TIA at the output of the balance photodiode. Those phase shifts are achieved by the application of a current through A_i.

In the case where the coupling losses are negligible, $a \approx 1$, the relationship between the add-drop through and drop transfer functions is $T_p = T_d - 1$. In addition, if we connect the through and drop ports into a balanced photodiode and TIA as in Fig. 3(a), we get an effective transfer function of $g(T_p - T_d)$ where g is the gain of the TIA. Therefore, we get a modulation of:

$$I_{mod} = g(T_p(\phi_i) - T_n(\phi_i))|E_0|^2 = A_i. \tag{12}$$

At the output of the balanced photodiode, the transfer function of $T_p - T_d$ is shown by the blue triangle curve in Fig. 3(b). Orange circle and green curves are Lorentzian line shapes, centered in the initial phase where MRR is in resonance with the incoming light, described by Eqs. (11) and (10), respectively. Differently, Fig. 3(c) and (d), are centered in a modified phase ($\phi + 0.2$), according to a specific value of the current A_i. Here we aim to demonstrate how to represent positive and negative kernel values in analog photonics. This can be achieved by incorporating a balanced-PD at the output of the add-drop MRR. In panels (c) and (d), the blue curves show such positive and negative kernel values from the drop and the through outputs, respectively. The orange triangle curves show the TIA transfer function $g(T_p - T_d)$, where g amplifies $T_p - T_d$ by a factor of two.

II.3. Dot Products with Photonics

The fundamental operation of a convolution is the dot product of two vectorized matrices. Therefore, one needs to understand how to compute a vector dot product using photonics before proposing an architecture capable of performing convolutions.

A wavelength multiplexed signal consists of k electromagnetic waves, each with angular frequency ω_i, $i = 1, \ldots, k$. If it is assumed that each wave has an amplitude of E_0, a power enveloping function μ_i whose modulation frequency is significantly smaller than ω_i, then the slowly varying envelope approximation and a short-time Fourier transform can be used to derive an expression for the multiplexed signal in the frequency domain:

$$E_{mux}(\omega) = \sum_{i=1}^{k} E_0 \sqrt{\mu_i}\delta(\omega - \omega_i), \tag{13}$$

where $\delta(\omega - \omega_i)$ is the Dirac delta function and $\mu_i \geq 0$, since power envelopes are not negative. If the enveloping function is prevented from amplifying the electric field, μ_i can further be restricted to the domain $0 \leq \mu_i \leq 1$. Next, we introduce tunable linear filters $H^+(\omega)$ and $H^-(\omega)$ such that when they interact with multiple fields, the following weighted signals are created:

$$\begin{aligned} E_w^-(\omega) &= H^-(\omega)E_{mux}(\omega), \\ E_w^+(\omega) &= H^+(\omega)E_{mux}(\omega). \end{aligned} \tag{14}$$

Assuming that the two signals are fed into a balanced photodiode (balanced PD) with spectral response $R(\omega)$, the induced photocurrent is described by:

$$\begin{aligned} i_{PD} &= \int_{-\infty}^{\infty} d\omega R(\omega) \left(\left|E_w^+(\omega)\right|^2 - \left|E_w^-(\omega)\right|^2 \right), \\ &= \int_{-\infty}^{\infty} d\omega R(\omega) \left(\left|H^+(\omega)\right|^2 - \left|H^-(\omega)\right|^2 \right) |E_{mux}(\omega)|^2, \\ &= \sum_{i=0}^{k-1} R(\omega_i) \left(\left|H^+(\omega_i)\right|^2 - \left|H^-(\omega_i)\right|^2 \right) E_0 r_i. \end{aligned} \tag{15}$$

Assuming that $R(\omega)$ is roughly constant in the area of spectral interest, one can set $A_i = E_0 R_0 \mu_i$ and $F_i^* = |H^+(\omega_i)|^2 - |H^-(\omega_i)|^2$ resulting in a photocurrent equal to

$$i_{PD} = \sum_{i=1}^{k} A_i F_i^* = \vec{A} \cdot \vec{F}^*. \tag{16}$$

The through and drop ports of a MRR can be used to implement the linear filters H^+ and H^- such that $|H^+|^2 = T_d$ and $|H^-|^2 = T_d$. Knowing that $T_p = T_d - 1$ with minimal losses, we can set a particular weight using:

$$F_i^* = 2T_d(\phi_i) - 1, \tag{17}$$

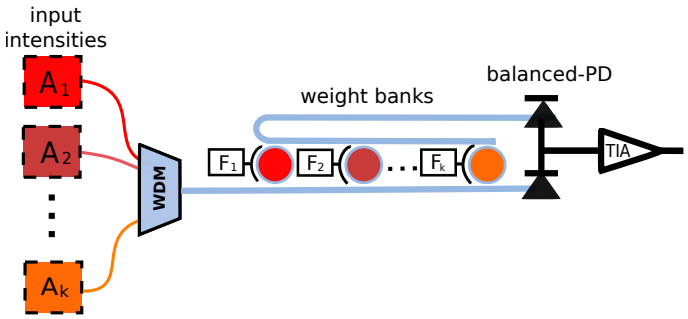

Figure 4. An electro-optic architecture that performs dot products. A_i $(i = 1, \ldots, k)$ are input elements encoded in intensities, multiplexed by a WDM and linked to the weight banks via a silicon waveguide. F_i are filter values that modulate the MRRs in the PWB. Drop and through output ports are connected to a balanceD-PD, where the matrix multiplication is performed, followed by an amplifier TIA.

where the phase, ϕ_i can obtained by using Eq. 10 and Eq. 11 to get:

$$\phi_i = \arccos\left[-\frac{1}{2r^2a}\left(\frac{2(1-r)^2a}{F_i^*+1} - 1 - (r^2a)^2\right)\right], \quad (18)$$

we can see that F_i^* can be between -1 and 1. Since T_d is a filter that only represents values between 0 and 1. In order to perform a dot product with a weight vector $\vec{w}$ whose components are not limited to the range -1 to 1, a gain g_{TIA} can be applied to the photocurrent such that:

$$\begin{aligned}\vec{A}\cdot\vec{F} &= g_{TIA}\vec{A}\cdot\vec{F}^* \\ &= g_{TIA}\sum_{i=1}^{k} A_i F_i^*,\end{aligned} \quad (19)$$

if:

$$g_{TIA} = \max_{1\le i\le k} |F_i|, \quad (20)$$

then,

$$\vec{F} = g_{TIA}\vec{F}^*; \quad (21)$$

assuming that each ϕ_i corresponds to a weighting of w_i^*. This electronic gain can be performed using a transimpedance amplifier (TIA), which can be manufactured in a standard CMOS process [24] and packaged or integrated with the photonic chip [21]. A diagram of the electro-optic architecture described in this section is presented in Fig. 4. From now on, this amalgamation of electronic and optical components is referred as a photonic weight bank (PWB). PWBs similar to the one in Fig. 4 have been successfully implemented in the past [11, 25, 26].

We can represent negative inputs between -1 and 1 by modifying the power enveloping function to $\mu_i = \frac{1}{2}(x_i + 1)$. If the same set of derivations is followed, we can modify Eq. (21) to be:

$$\vec{x}\cdot\vec{w} = g\left(\sum_{i=1}^{k} A_i F_i^* + \sum_{i=1}^{k} E_0 R_0 F_i^*\right). \quad (22)$$

The second term in this sum is a predictable bias current term that conceptually be subtracted before feeding into the TIA. This is a disadvantage of supporting negative inputs, as additional optical or electronic control circuitry would need to be designed. Another trade-off is a loss in precision due to a larger range of inputs needing to be represented, analogous to the loss in precision with signed integers for classical computing.

III. PERFORMING CONVOLUTIONS USING PHOTONICS

The goal of this section is to present a photonic architecture capable of performing convolutions for CNNs. This new architecture is called DEAP.

For a maximum number of input channels D_m and a maximum kernel edge length R_m as bounding parameters for DEAP, we represent the range of convolutional parameters that a particular implementation of DEAP can support. If a convolutional parameter described in Table (I) does not have a complementary bounding parameter, it means that the DEAP architecture can support for arbitrary values of said convolutional parameter.

III.1. Producing a Single Convolved Pixel

First, we consider an architecture that can produce one convolved pixel at a time. To handle convolutions for kernels with dimensionality up to $R_m \times R_m \times D_m$, we will require R_m^2 lasers with unique wavelengths since a particular convolved pixel can be represented as the dot product of two $1 \times R_m^2$ vectors. To represent the values of each pixel, we require $D_m R_m^2$ modulators (one per kernel value) where each modulator keeps the intensity of the corresponding carrier wave proportional to the normalized input pixel value. The R_m^2 lasers are multiplexed together using wavelength division multiplexing (WDM), which is then split into D_m separate lines. On every line, there are R_m^2 all-pass MRRs, resulting in $D_m R_m^2$ MRRs in total. Each WDM line will modulate the signals corresponding to a subset of R_m^2 pixels on channel k, meaning that the modulated wavelengths on a particular line correspond to the pixel inputs $(A_{m,n,k})_{n\in[j,j+R_m]}^{m\in[i,i+R_m]}$ where $k \in [1, D_m]$.

The D_m WDM lines will then be fed into an array of D_m PWBs. Each PWB will contain R_m MRRs with the weights corresponding to the kernel values at a particular channel. For example, the PWB on line k should contain the vectorized weights for the kernel $(F_{m,n,k})_{n\in[1,R_m^2]}^{m\in[1,R_m^2]}$. Each MRR within a PWB should be

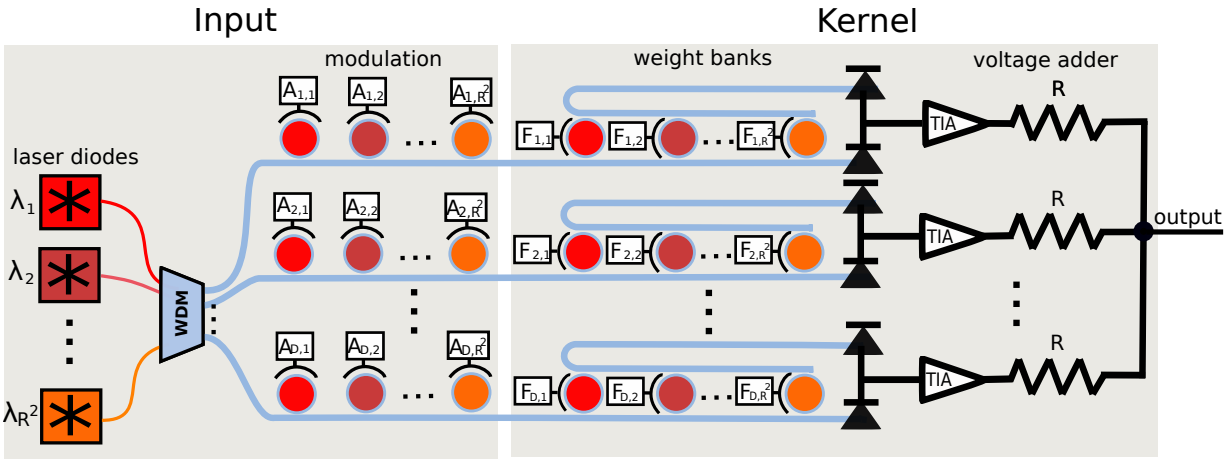

Figure 5. Photonic architecture for producing a single convolved pixel. Input images are encoded in intensities $A_{l,k}$, where the pixel inputs $A_{m,n,k}$ with $m \in [i, i+R_m], n \in [j, j+R_m], k \in [1, D_m]$ are represented as $A_{l,h}$, $l = 1, \ldots, D$ and $h = 1, \ldots, R^2$. Considering the boundary parameters, we set $D = D_m$ and $R = R_m$. Likewise, the filter values $F_{m,n,k}$ are represented as are represented as $F_{l,h}$ under the same conditions. We use an array of R^2 lasers with different wavelengths λ_h to feed the MRRs. The input and kernel values, $A_{l,h}$ and $F_{l,h}$ modulate the MRRs via electrical currents proportional to those values. Once the matrix parallel multiplications are performed, the voltage adder has the function to add all signals from weight banks. Here, **R** are resistance values. Then the output is the convolved feature.

tuned unique the resonant wavelength within the multiplexed signal. The outputs of the weight bank array are electrical signals, each proportional to the dot product $(F_{m,n,k})_{n\in[1,R_m^2]}^{m[1,R_m^2]} \cdot (A_{p,q,k})_{q\in[j,j+R_m^2]}^{p\in[i,i+R_m^2]}$. Finally, the signals from the weight banks need to be added together. This can be achieved using a passive voltage adder. The output from this adder will therefore be the value of a single convolved pixel. Fig. 5 shows a complete picture of what such an architecture would look like.

To perform a convolution with a kernel edge length less than R_m, one can set $(F_{m,n,k})_{n\in[R+1,R_m]}^{m\in[R+1,R_m]}$ to zero. Similarly, if the dimensionality of the kernel is less than D_m, then the modulators $(A_{m,n,k})_{n\in[1,W]}^{m\in[1,H]}$ should also be set to zero, with $k \in [D+1, D_m]$ in this case.

Algorithm 1 Convolutions for CNNs using DEAP

```
A is the input image
F is the kernel
R is the edge length of the kernel
O is a memory block to store the convolution
S is the stride
H and W are the height and width of the input image
function CONVOLVE(A, F, R, O, S, H, W)
    for (k = 1; k ≤ K; k = k + 1) do
        load kernel weights from F[:,:,:,k]
        for (h = 1; h ≤ H − R + 1; h = h + S) do
            for (w = 1; w ≤ W − R + 1; w = w + S) do
                load inputs from A[h:min(h+T,H), w:min(w+R,W),:]
                perform convolution
                store results in O[h/S,w/S,k]
            end for
        end for
    end for
end function
```

III.2. Performing a Full Convolution

In the previous section, we have discussed how DEAP can produce a single convolved pixel. In order to perform a convolution of arbitrary size, one would need to stride along the input image and readjust the modulation array. Since the same kernel is applied across the set of inputs, the weight banks do not need to be modified until a new kernel is applied. Fig. 6(a) demonstrates this process on an input with $S = 1$. To handle $S \geq 1$, the inputs being passed in to DEAP should also be strode accordingly. In this approach, the inputs should have been zero padded before being passed into DEAP. In pseudocode, performing a convolution with K filters can be implemented as shown in Algorithm 1.

The DEAP architecture also allows for parallelization by treating the photonic architecture proposed in the previous section as a single output "convolutional unit". However, by creating n_{conv} instances of these convolutional units, you could produce n_{conv} pixels per cycle by passing in the next set of inputs per unit. This is demonstrated in Fig. 6(b) for $n_{conv} = 2$. The computation of output pixels can be distributed across each convolutional unit, resulting in a runtime complexity of $O\left(\frac{KHW}{S^2 n_{conv}}\right)$.

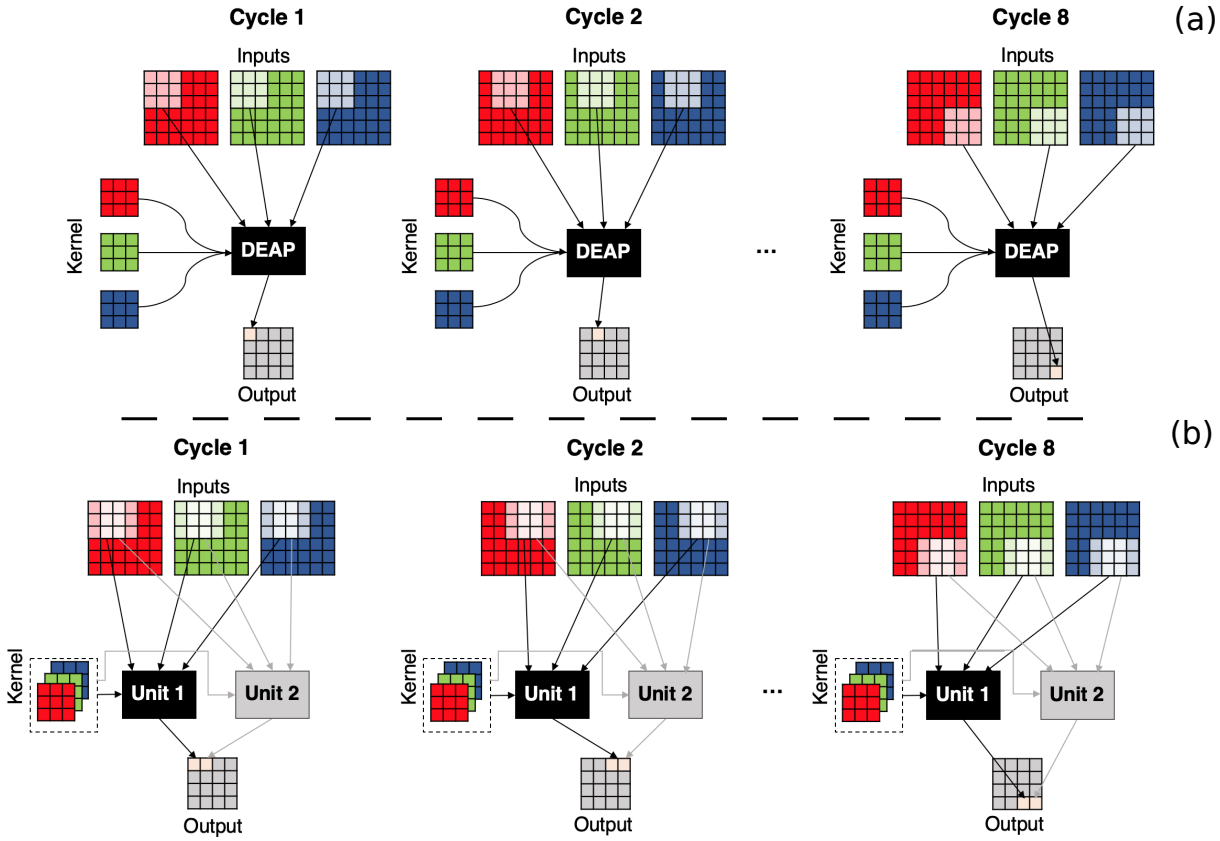

Figure 6. (a) Cycling through a convolution using DEAP. (b) Performing a convolution with two convolutional units.

IV. PHOTONIC CONVOLUTIONAL NEURAL NETWORKS

In this section, we show how DEAP can be used to run a CNN. CNNs are a type of neural network that were developed for image recognition tasks. A CNN consists of some combination of convolutional, nonlinear, pooling and fully connected layers [27], see Fig. 7(a). As introduced previously, convolutions perform a highly efficient and parallel matrix multiplication using kernels [3]. Furthermore, since kernels are typically smaller than the input images, the feature extraction operation allows efficient edge detection, therefore reducing the amount of memory required to store those features.

CNNs are networks suitable to be implemented in photonic hardware since they demand fewer resources to do matrix multiplication and memory usage. The linear operation performed by convolutions allows single feature extraction per kernel. Hence, many kernels are required to extract as many features as possible. For this reason, kernels are usually applied in blocks, allowing the network to extract many different features all at once and in parallel.

In feed-forward networks, it is typical to use a rectified linear unit (ReLU) activation function. Since ReLUs are linear piecewise functions that model an overall nonlinearity, they allow CNNs to be easily optimized during training. The pooling layer introduces an stage where a set of neighbor pixels are encompassed in a single operation. Typically, such operation consists in the application of a function that determines the maximum value among neighboring values. An average operation can be implemented likewise. Both approaches describe max and average pools, respectively. This statistical operation allows for a direct down-sampling of the image, since the dimensions of the object are reduced by a factor of two. From this step, we aim to make our network invariant and robust to small translations of the detected features.

The triplet, convolution-activation-pooling, is usually repeated several times for different kernels, keeping invariant the pooling and activation functions. Once all possible features are detected, the addition of a fully connected layer is required for the classification stage. This layer prepares and shows the solutions of the task.

CNNs are trained by changing the values of the kernels, analogous to how feed-forward neural networks are trained by changing the weighted connections [28]. The estimated kernel and weight values are required in the testing stage. In this work, this stage is performed by our on-chip DEAP CNN. Figure 7(b) shows a high-level overview of the proposed testing on-chip architecture. Here, the testing input values stored in the PC modulate the intensities of a group of lasers with identical powers but unique wavelengths. These modulated inputs would

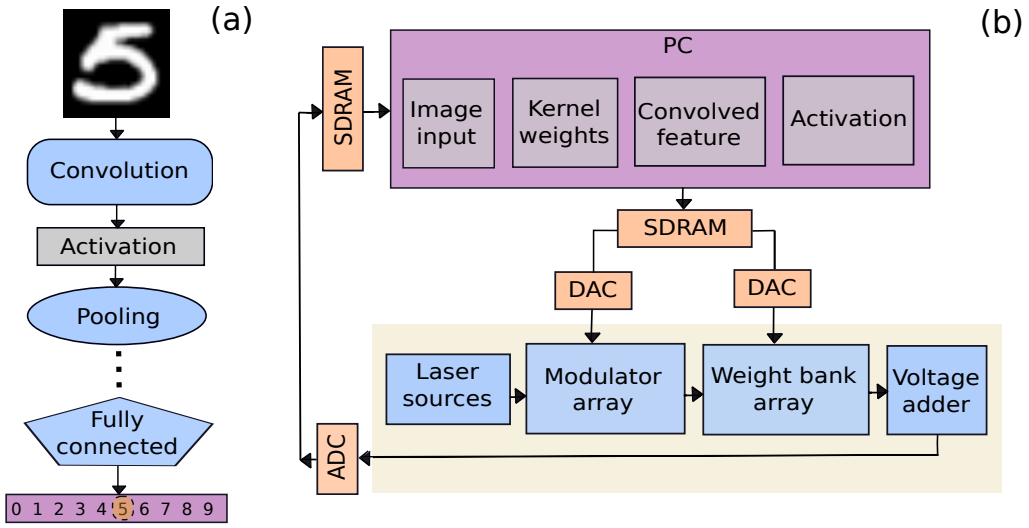

Figure 7. Block diagrams that describe: (a) a typical CNN, which contains convolutions, activation functions, pooling and fully connected layers. In this case we exemplify such diagram using MNIST-based recognition task that predicts the number 5; and (b) the DEAP architecture. In the computer (PC) the input image, kernel weights, and convolved features are stored. Also, the commands to implement the activation function off-chip are stored in the PC. The input image, kernel weights and convolved features are transferred to the chip via DACs from the SDRAM. Then, the convolution is performed on-chip. Finally the output is digitalized via an ADC and stored in a SDRAM connected to the computer.

be sent into an array of photonic weight banks, which would then perform the convolution for each channel. The kernels obtained in the training step are used to modulate these weight banks. Finally, the outputs of the weight banks would be summed using a voltage adder, which produces the convolved feature. This simulator works using the transfer function of the MRRs, through port and drop port summing equations at the balanced PDs, and the TIA gain term to simulate a convolution. The simulator assumes that the MRRs can only be controlled with $7-$bits of precision as that has been empirically observed in a lab setting. The MRR self-coupling coefficient is equal to the loss, $r = a = 0.99$[29] in Eqs. (6) (10) and (11).

The interfacing of optical components with electronics would be facilitated by the use of digital-to-analog converters (DACs) and analog-to-digital converters (ADCs), while the storage of output and retrieving of inputs would be achieved by using memories GDDR SDRAM. The SDRAM is connected to a computer, where the information is already in a digital representation. Then, the implementation of the ReLU nonlinearity and the reuse of the convolved feature to perform the next convolution can be performed. The idea is to use the same architecture to implement the triplet convolution-activation-pooling on hardware.

In this work, we trained the CNN to perform image recognition on the MNIST dataset. The training stage uses the ADAM optimizer and back-propagation algorithm to compute the gradient function. The optimized parameters to solve MNIST can be categorized in two groups: (i) two $5\times5\times8$ different kernels and (ii) two fully connected layers of dimensions 128×800 and 128×10; and their respective bias terms. These kernels are then defined by eight 5×5 different filters. In the following we use our DEAP CNN simulator to recognize new input images, obtained from a set of 500 images, which are intended to be used for the test step. Our simulator only works at the transfer level and does not simulate noise or distortion from analog components. The process of feature extraction performed by the DEAP CNN is illustrated in Fig. 8(a). As it can be seen in the illustration, a 28×28 input image from the test dataset is filtered by a first $5 \times 5 \times 8$ kernel, using stride one. The output of this process is a $24 \times 24 \times 8$ convolved feature, with a ReLU activation function already applied. Following the same process, the second group of filters is applied to the convolved feature to generate the second output, i.e. a $20 \times 20 \times 8$ convolved feature.

After the second ReLU is applied to the output, average pooling is utilized for invariance and down-sampling of the convolved features. The average pooling is implemented by a 2×2 kernel whose elements are all $1/4$. However, the stride one was kept; therefore the pooled feature has dimensionality $19\times19\times8$. The down-sampling is implemented offline: from the $19 \times 19 \times 8$ output, a simple algorithm extracts the elements that have even indexes. The result of this process is a $10 \times 10 \times 8$ pooled output. Finally, the first fully connected layer is fed through by the flattened version of the pooled object. The resultant vector feeds the last fully connected layer, where the result of the MNIST classification appears.

The results of the MNIST task solved by our simulated DEAP CNN is shown by Fig. 8(b). For a test set of 500 images, we obtained an overall accuracy of 98%. This performance was compared to the results obtained using a standard two-layers CNN including a max pooling layer. We found that This standard network achieves

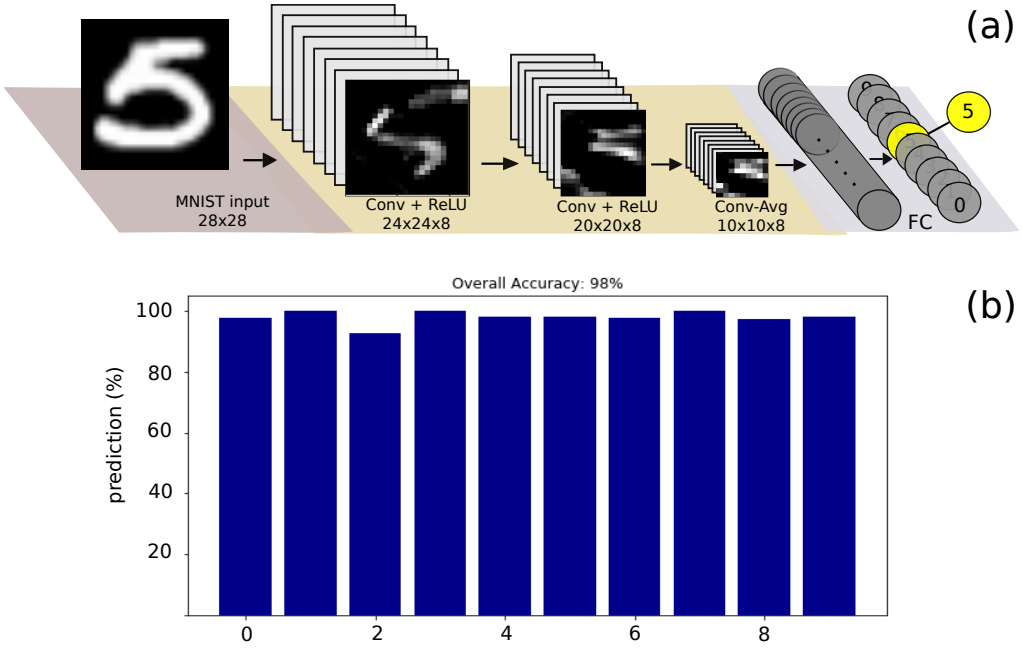

Figure 8. (a) An illustrative block diagram of the two-layers DEAP CNN solving MNIST. (b) Results of the MNIST task using a simulated DEAP CNN.

an overall accuracy of 98.6%. Therefore, we can conclude that our simulator is sufficiently robust despite the 7−bits of precision considered in the DEAP CNN simulation.

V. ENERGY AND SPEED ANALYSES

V.1. Energy Estimation

The energy used by a single DEAP convolutional unit depends on the R and D parameters. The 100-wavelength limitation for MRRs constrains the maximum R to be 10, as each multiplexed waveguide will store R^2 signals. The number of MRRs used in the modulator array is equal to R^2D, meaning that only certain D and R^2 values are allowed for a finite number of MRRs. Assuming that a maximum of 1024 MRRs can be manufactured in the modulator array, a convolutional unit can support a large kernel size with a limited number of channels, $R = 10$, $D = 12$, or a small kernel size with a large number of channels, $R = 3$, $D = 113$. We will consider both edge cases to get a range of energy consumption values. For the smaller convolution size, we will have R^2 lasers, R^2 MRRs and DACs in the modulator array, R^2D MRRs and D TIAs in the weight bank array and one ADC to convert back into digital signal. With 100 mW per laser, 19.5 mW per MRR, 26 mW per DAC, 17 mW per TIA [30] and 76 mW per ADC, we get an energy usage of 112 W for the large kernel size and 95W for the smaller kernel size. Therefore, we estimate a single convolution unit to use around 100 W when 1024 modulators are used to represent inputs.

V.2. DEAP Performance

The time it takes for light to propagate from the WDM to before the balanced PDs is estimated by the following equation:

$$t_{prop} = \frac{k2\pi r_{MRR}}{c} \tag{23}$$

where c is the speed of light $2\pi r_{MRR}$ is the circumference of the MRR and k is the number of MRRs. Assuming 100 MRRs with a radius of 10 m [11, 31], the PWB gets a propagation time of around 21 ps and a throughput of $1/t_{prop} = 50$ GS/s. The bottlenecks come from the fact that the balanced PDs has a throughput of 25 GS/s[30] and the TIA has a throughput of 10 GS/s[32]. An individual MRR can be modulated at speeds of 128 GS/s[31], meaning that the modulation frequency of the MRRs does not bottleneck the throughput of the PWB.

The throughput of a PWB is around 5 GS/s. The DACs[33] and ADCs[34] both operate at 5 GS/s and support to 7-bits. The GDDR6 SDRAM operates at 16 G with a 256-bit bus size[35]. Consequently, the speed of the system is limited by the throughput of the DACs/ADCs, resulting in DEAP producing a single convolved pixel at 5 GS/s or $t = 200$ ps.

DeepBench [17] is an empirical dataset that contains how long various types of GPUs took to perform a convolution for a given set of convolutional parameters. Table (II) contains the parameters used for each of these benchmarks, and Table (III) contains the power consumption.

The speeds of various GPUs were directly taken from Ref. [17], while the speed of the convolution was esti-

Table II. Benchmarking parameters for DEAP

W	H	D	N	K	R_w	R_h	S
700	161	1	4	32	5	20	2
112	112	64	8	128	3	3	1
7	7	832	16	256	1	1	1

Table III. Benchmarked GPUs with power consumption

GPU	Power Usage (W)
AMD Vega FE[36]	375
AMD MI25[37]	300
NVIDIA Tesla P100[38]	250
NVIDIA GTX 1080 Ti[39]	250

mated using the following equation:

$$t_{runtime} = 200 \text{ ps} \times \frac{NK}{n_{conv}} \left(\frac{H-R}{S} + 1 \right) \left(\frac{W-R}{S} + 1 \right). \tag{24}$$

In some of the benchmarks, the kernels edge lengths were not equal, hence the parameters R_w and R_h which correspond to the width and height of the kernels. For each of the selected benchmarks, the parameters $R^2 D \leq 1024$, meaning that the convolutional network is compatible with DEAP implementations.

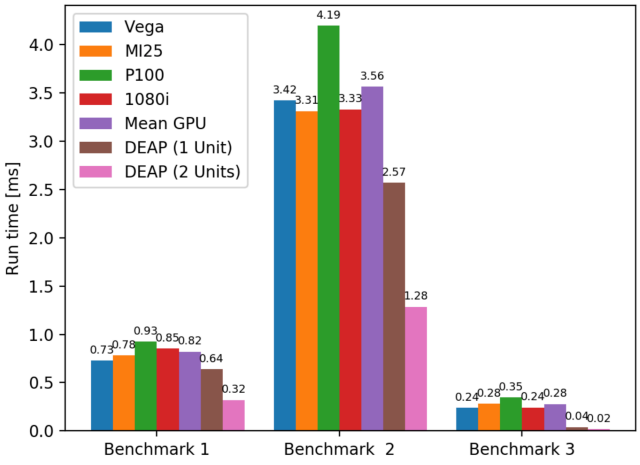

Figure 9. Estimated DEAP convolutional runtime compared to actual GPU runtimes from DeepBench benchmarks

The estimated DEAP runtimes using one and two convolutional units were plotted against actual DeepBench runtimes in Fig. 9. From this, we can see that using two convolutional units performs slightly better than all the GPU benchmarks. While mean GPUs power consumption is 295 W, DEAP with a single convolutional unit sues about 110 W. Therefore, DEAP can perform convolutions between 1.4 and 7.0 faster than the mean GPU runtime while using 0.37 times the energy consumption. Using two convolutional units doubles the speed of DEAP, meaning that DEAP can be between 2.8 and 14 times faster than a conventional GPU while using almost 0.75 times the energy consumption. DEAP with a single unit performing at a speed somewhat similar to the GPUs is expected.

VI. CONCLUSION

We have proposed a photonic network, DEAP, suited for convolutional neural networks. DEAP was estimated to perform convolutions between 2.8 and 14 times faster than a GPU while roughly using 0.75 times the energy consumption. A linear increase in processing speeds corresponds to a linear increase in energy consumption, allowing for DEAP to be as scalable as electronics.

High level software simulations have shown that DEAP is theoretically capable of performing a convolution. We demonstrate that our DEAP CNN is capable of solving MNIST handwritten recognition task with an overall accuracy of 98%. The largest bottlenecks is the I/O interfacing with digital systems via DACs and ADCs. If photonic DACs[40] and ADCs[41] are to be built with higher bit-precisions, the speedup over GPUs could be even higher. If higher bit precision photonic DACs and ADCs are able to be built, replacing the electronic components with optical ones can significantly decrease the runtime.

In order to realize a physical implementation, there are a number of issues that still need to be solved. Packaging a silicon photonic with an electronic chip with high I/O count is a challenging RF engineering task, but it is a central thrust in the roadmap for silicon photonic foundries [21]. There also needs to be control circuitry that routes the outputs of the SDRAM into the relevant DACs and from the ADCs into the SDRAM. Since we assume that the control circuitry can operate significantly faster than a memory access, we believe it will have a negligible impact on the overall throughput. Another issue is that DEAP processes data in the analog domain, whereas GPUs perform floating point arithmetic. Though floating-point arithmetic does have some degree of error due to rounding in the mantissa, their errors are deterministic and predictable. On the other hand, the errors from photonics are due to stochastic shot, spectral, Johnson-Nyquist and flicker noises, as well as quantization noise in the ADC, and distortion from the RF signals applied to the modulators. However, artificially adding random noise to CNNs have been shown to reduce over-fitting [42], meaning that some degree of stochastic behaviour is tolerable in the domain of machine learning problems.

Finally, MRRs have only been shown to have up to 7-bits of precision, which is significantly smaller than the range precision supported by even half-precision (16-bit) floating point representations. In conclusion, photonics has the potential to perform convolutions at speeds faster

than top-of- the-line GPUs while having a lower energy consumption. Moving forward, the greatest challenges to overcome have to do with increasing the precision of photonic components so that they are comparable to classical floating-point representations. Overall, silicon photonics has the potential to outperform conventional electronic hardware for convolutions while having the ability to scale up in the future.

ACKNOWLEDGMENT

Funding for B.J.S., B.A.M., H.B.M., and V.B. was provided by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Queen's Research Initiation Grant (RIG).

[1] X. Li, G. Zhang, H. H. Huang, Z. Wang, and W. Zheng, in *2016 45th International Conference on Parallel Processing (ICPP)* (2016) pp. 67–76.
[2] M. Jaderberg, A. Vedaldi, and A. Zisserman, CoRR **abs/1405.3866** (2014), arXiv:1405.3866.
[3] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (The MIT Press, 2016).
[4] Y. LeCun and C. Cortes, (2010).
[5] F. Duport, A. Smerieri, A. Akrout, M. Haelterman, and S. Massar, Scientific Reports **6**, 22381 EP (2016).
[6] D. Brunner, M. C. Soriano, C. R. Mirasso, and I. Fischer, Nature Communications **4**, 1364 (2013).
[7] K. Vandoorne, P. Mechet, T. Van Vaerenbergh, M. Fiers, G. Morthier, D. Verstraeten, B. Schrauwen, J. Dambre, and P. Bienstman, Nature Communications **5**, 3541 EP (2014).
[8] L. Larger, M. C. Soriano, D. Brunner, L. Appeltant, J. M. Gutierrez, L. Pesquera, C. R. Mirasso, and I. Fischer, Optics Express **20**, 3241 (2012).
[9] P. R. Prucnal and B. J. Shastri, *Neuromorphic Photonics* (CRC Press, Taylor & Francis Group, Boca Raton, FL, USA, 2017).
[10] A. N. Tait, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, Journal of Lightwave Technology **32**, 4029 (2014).
[11] A. N. Tait, A. X. Wu, T. F. de Lima, E. Zhou, B. J. Shastri, M. A. Nahmias, and P. R. Prucnal, IEEE Journal of Selected Topics in Quantum Electronics **22**, 312 (2016).
[12] A. N. Tait, T. Ferreira de Lima, M. A. Nahmias, H. B. Miller, H.-T. Peng, B. J. Shastri, and P. R. Prucnal, arXiv e-prints , arXiv:1812.11898 (2018), arXiv:1812.11898 [physics.app-ph].
[13] T. W. Hughes, M. Minkov, Y. Shi, and S. Fan, Optica **5**, 864 (2018).
[14] Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and M. Soljačić, Nat. Photonics **11**, 441 (2017).
[15] T. F. de Lima, H. Peng, A. N. Tait, M. A. Nahmias, H. B. Miller, B. J. Shastri, and P. R. Prucnal, Journal of Lightwave Technology **37**, 1515 (2019).
[16] A. Mehrabian, Y. Al-Kabani, V. J. Sorger, and T. A. El-Ghazawi, CoRR **abs/1807.08792** (2018), arXiv:1807.08792.
[17] B. Research, Deepbench.
[18] V. Bangari, B. Marquez, H. Miller, and B. J. Shastri, DEAP, `https://github.com/Shastri-Lab/DEAP` (2019).
[19] G. Tan, L. Li, S. Triechle, E. Phillips, Y. Bao, and N. Sun, in *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '11 (ACM, New York, NY, USA, 2011) pp. 35:1–35:11.
[20] S. Chetlur, C. Woolley, P. Vandermersch, J. Cohen, J. Tran, B. Catanzaro, and E. Shelhamer, CoRR **abs/1410.0759** (2014), arXiv:1410.0759.
[21] A. Rahim, T. Spuesens, R. Baets, and W. Bogaerts, Proceedings of the IEEE **106**, 2313 (2018).
[22] M. A. Nahmias, B. J. Shastri, A. N. Tait, T. F. de Lima, and P. R. Prucnal, Opt. Photon. News **29**, 34 (2018).
[23] W. Bogaerts, P. De Heyn, T. Van Vaerenbergh, K. De Vos, S. Kumar Selvaraja, T. Claes, P. Dumon, P. Bienstman, D. Van Thourhout, and R. Baets, Laser & Photonics Reviews **6**, 47 (2012), arXiv:arXiv:1208.0765v1.
[24] H. Zheng, R. Ma, and Z. Zhu, Analog Integrated Circuits and Signal Processing **90**, 217 (2017).
[25] M. Lipson, Journal of Lightwave Technology **23**, 4222 (2005).
[26] A. N. Tait, H. Jayatilleka, T. F. D. Lima, P. Y. Ma, M. A. Nahmias, B. J. Shastri, S. Shekhar, L. Chrostowski, and P. R. Prucnal, Opt. Express **26**, 26422 (2018).
[27] K. O'Shea and R. Nash, CoRR **abs/1511.08458** (2015), arXiv:1511.08458.
[28] K. Mehrotra, C. K. Mohan, and S. Ranka, *Elements of Artificial Neural Networks* (MIT Press, Cambridge, MA, USA, 1997).
[29] Y. Tan and D. Dai, Journal of Optics **20**, 054004 (2018).
[30] Z. Huang, C. Li, D. Liang, K. Yu, C. Santori, M. Fiorentino, W. Sorin, S. Palermo, and R. G. Beausoleil, Optica **3**, 793 (2016).
[31] J. Sun, R. Kumar, M. Sakib, J. B. Driscoll, H. Jayatilleka, and H. Rong, Journal of Lightwave Technology **37**, 110 (2019).
[32] M. Atef and H. Zimmermann, Analog Integr. Circuits Signal Process. **76**, 367 (2013).
[33] B. Sedighi, M. Khafaji, and J. C. Scheytt, International Journal of Microwave and Wireless Technologies **4**, 275 (2012).
[34] J. Fang, S. Thirunakkarasu, X. Yu, F. Silva-Rivas, C. Zhang, F. Singor, and J. Abraham, IEEE Transactions on Circuits and Systems I: Regular Papers **64**, 1673 (2017).
[35] I. Micron Technology, Gddr6 sgram mt61k256m32 8gb: 2 channels x16/x8 gddr6 sgram.
[36] I. Advanced Micro Devices, Radeon vega frontier edition (liquid-cooled) ().
[37] I. Advanced Micro Devices, Radeon instinct mi25 accelerator ().
[38] N. Corporation, Vidia tesla p100 gpu accelerator ().
[39] N. Corporation, Geforce gtx 1080 ti ().
[40] F. Zhang, B. Gao, X. Ge, and S. Pan, Optical Engineering **55**, 031115 (2015).
[41] M. A. Piqueras, P. Villalba, J. Puche, and J. Martí, in *2011 IEEE International Conference on Microwaves,*

Communications, Antennas and Electronic Systems (COMCAS 2011) (2011) pp. 1–6.

[42] Z. You, J. Ye, K. Li, and P. Wang, CoRR **abs/1805.08000** (2018), arXiv:1805.08000.