Research article

Moustafa Ahmed, Yas Al-Hadeethi, Ahmed Bakry, Hamed Dalir and Volker J. Sorger*

Integrated photonic FFT for photonic tensor operations towards efficient and high-speed neural networks

https://doi.org/10.1515/nanoph-2020-0055
Received January 25, 2020; accepted May 18, 2020; published online June 26, 2020

Abstract: The technologically-relevant task of feature extraction from data performed in deep-learning systems is routinely accomplished as repeated fast Fourier transforms (FFT) electronically in prevalent domain-specific architectures such as in graphics processing units (GPU). However, electronics systems are limited with respect to power dissipation and delay, due to wire-charging challenges related to interconnect capacitance. Here we present a silicon photonics-based architecture for convolutional neural networks that harnesses the phase property of light to perform FFTs efficiently by executing the convolution as a multiplication in the Fourier-domain. The algorithmic executing time is determined by the time-of-flight of the signal through this photonic reconfigurable passive FFT 'filter' circuit and is on the order of 10's of picosecond short. A sensitivity analysis shows that this optical processor must be thermally phase stabilized corresponding to a few degrees. Furthermore, we find that for a small sample number, the obtainable number of convolutions per {time, power, and chip area) outperforms GPUs by about two orders of magnitude. Lastly, we show that, conceptually, the optical FFT and convolution-processing performance is indeed directly linked to optoelectronic device-level, and improvements in plasmonics, metamaterials or nanophotonics are fueling next generation densely interconnected intelligent photonic circuits with relevance for edge-computing 5G networks by processing tensor operations optically.

Keywords: integrated photonic; metasurface; neural networks; optical convolutions.

***Corresponding author: Volker J. Sorger**, Department of Electrical and Computer Engineering, George Washington University, 20052, Washington, D.C., USA, E-mail: sorger@gwu.edu
Moustafa Ahmed, Yas Al-Hadeethi and Ahmed Bakry: Department of Physics, Faculty of Science, King Abdulaziz University, 21589, Jeddah, Saudi Arabia
Hamed Dalir: Omega Optics, Inc. 8500 Shoal Creek Blvd., 78757, Austin, Texas, USA

1 Introduction

In this post-Moore era the trend in signal processing and computing towards a higher degree of compute-system heterogeneity, specialized and domain-specific processors are gaining interest [1] such as exemplary GPU's augmenting CPU systems, or Tensor core Processor Units (TPU) outperforming GPU's on specific tasks [2]. A second macroscopic trend in information processing is driven by demand for machine learning (ML) tasks, specifically by deep-learning (DL) [3] enabled by large available data-sets [4]. The hardware for the ML tasks are neuromorphic inspired systems [5, 6], namely artificial neural networks (NN) consisting of the following general parts; (i) convolutional filter performed as weighted-additions (multiply-accumulate, MAC operations), (ii) nonlinear activation function ('threshold'), (iii) 'signal clean-up' often performed as a down sampling max-pooling or averaging, followed by (iv) a fully-connected layer. However, the majority (can approach 80%) of the processing overhead in DL are convolution-operations, which are a specific sub-task of generalized tensor processing (Figure. 1a). The latter includes all permutations of a two-part operation consisting of either part being a scalar, vector, or a matrix with each data dimension being one, N or N^2, respectively. A two-dimensional convolution, thus, consists of matrix multiplications, where the first matrix $\boldsymbol{A_{i,j}}$ is the input data, and the second matrix the 'weighting' kernel $\boldsymbol{B_{i,j}}$, where the dimension of the matrix is i x j, however, for ensuring generality, i does not have to equal j, and the input matrix does not have to match the kernel matrix (Figure 1b). Note, the following introduction citations are to be taken exemplary in nature and are hence not-exhaustive, as this work is not considered a comprehensive review of the field.

This paper assumes the reader is familiar with the unique strengths and challenges of optics in regards to general information processing, and Ref [7] provides a summary; in brief strengths include: a) non-iterative $\boldsymbol{O}(1)$ processing in the analog domain, b) ~ps short delay when photonic integrated circuits are used, c) synergistic convolutions due to natural Fourier transformation (FT)

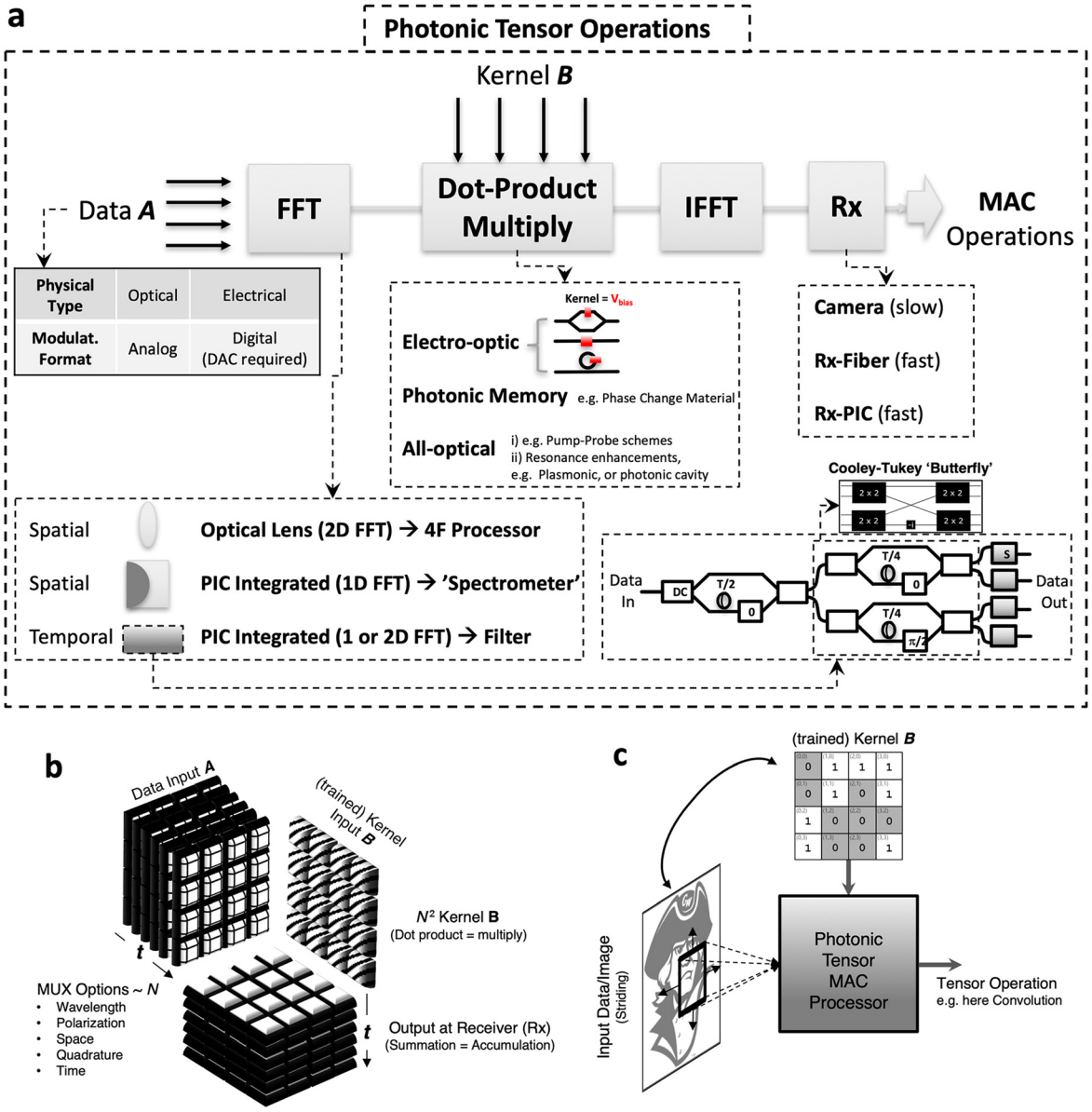

Figure 1: Photonic Tensor Processing paradigm utilizing Fourier-optics. a, Kernel data matrix B is dot-product multiplied with data matrix A either electro-optically, via a photonic nonvolatile memory, or all-optically in the Fourier domain. The Fourier transform (FT) can be performed spatially via a free-space or photonic integrated lens, or temporally via an Cooley-Turkey butterfly filter in integrated photonics, where 2×2 optical couplers provide addition and subtraction while a small path-length difference in one of the waveguide branches provides the reciprocal root-of-unity change in phase. A receiver (Rx) performed incoherently via a camera, fiber- or integrated photonics receiver arrays complete the multiply-accumulate (MAC) operation of the tensor processing unit. Depending on the dimensionality of the *A* and *B*, all permutations of tensor processing functions can be performed such as matrix-matrix, vector-matrix, scalar multiplication of vectors or matrices, convolutions etc. b, Exemplary 4×4 photonic tensor operation and scaling (*N*) options for dot-product engine and detection options from panel a. c, Within the category of tensor operations fall convolutions where the kernel *B* strides across the data input useful in convolutional neural networks (CNN), which constitutes the majority (~80%) of classification networks overhead.

performed by optical components (this is used in this work), d) no capacitive wire charging synergistic with non-Van Neumann distributed architectures such as NNs. Challenges of optics, however, include e) ensuring sufficient signal-to-noise (SNR) due to the analog nature, f) non-existence of nonvolatility (yet this can be introduced, for instance, via phase-change materials), g) weak and hence bulky electro-optic conversion, and h) system packing complexity arising from alignment requirements.

Harnessing the strengths of optics for emerging processors bears much potential to free circuits from charging wires, while utilizing massive parallelism paradigms [8].

From the top, optical approaches include free-space processors [9–12] and integrated photonics [13–16], which further divide in analog vs. spiking signaling [17–18]. Focusing next on the MAC operation to perform the convolution of CNNs, we turn to Figure 1, which shows a variety of design options; the optical data input can be either optical (e.g. from a lens) or electronic (e.g. from a sensor, or other electronic processor, or memory), and be modulated in an analog fashion or be in a digital format. In the latter a digital-to-analog converter (DAC) may be needed front-end, and an ADC back-end. Next, the dot-product multiplication to be performed as part of core of the tensor processor can be achieved in a multitude of options such as via electro-optic weighting by controlling the amplitude such as via an electro-absorption modulator [19, 20], and phase-shifter using an Mach Zehnder interferometer (MZI) [16, 21–23], or via photonic nonvolatile memory such as enabled by phase-change-material weighted states [24], or all-optically [25]. Each of these options can be performed non-resonantly (i.e. broadband) or resonantly, where the latter allows reducing the drive voltage, yet may require resonance stabilization which can be energy costly if thermal solutions are used. On that note, in general, the Purcell-factor improves electro-optic cavity-based device performance [26–28], hence also low mode volume options can be utilized.

While the focus of this paper is on the convolutional filtering via tensor operation (Figure 1a–c), the other mandatory NN functions can also be realized in optics and photonics, such as the nonlinear activation function [29, 30], max pooling [31], and fully connected layers [16]. Given that the above mentioned high-relevance of CNNs performing the convolution operations in the optical domain is of significant interest, next we turn to scaling laws of convolution processing. Since a convolution is a high dimensional $\sim N^3$ problem, parallelization strategies such as multiplexing is key, which hence is synergistic to optics and photonics [32–34]. An interesting inspiration can be borrowed from Fourier optics [9]; instead of performing the cumbersome (all-to-all) convolution between the data and the kernel, a simpler dot-product multiplication can be performed in the in the Fourier domain instead. While this is well known concept, it is not beneficial to perform this domain crossing (i.e. Fast Fourier transforms (FFT) and inverse FFT, IFFT) in electronics, since such transformations are rather costly. Instead here we select a direct approach; while complexity reduction algorithms exist reducing the $\boldsymbol{O}(N^3)$ complexity to $\boldsymbol{O}(N^{2x})$, these do not lead to a significant compute reduction. In contrast, in optics a multitude of passive FFT options exist (lower left, Figure. 1a), including spatial FFTs such as via a free-space lens leading to a 4F system configuration [9, 12, 36], an PIC-integrated star-coupler [31, 36] or 1D planar lens [37].

To gauge the possible performance potential of a 4F system, let us consider a lens with focal length of 15 mm and 15 mm diameter which has an optical latency of 0.16 ns on the marginal ray and at 1550 nm, 46.8 million diffraction limited pixels enable a throughput of 300 PHz. If each pixel is modulated with 8 bits, this results in 2.4 exabits/s of signal processing in a volume of 5.3 cm^3 [12]. Naturally, this is an upper limit of what may be physically possible and engineering challenges will place a much lower realistic bound. Nonetheless it stands to reason, that Fourier-optics based accelerators bear much potential for in signal processing and special purpose processors to perform correlations and convolutions as a fundamental building block. While free-space 4F systems are capable of high parallelism such as by utilizing digital light processing (DLP) technology [35, 38], they are limited with respect to a) delay given the often meter-scale setups, b) hence a burdoned by large and bulky footprints, and c) slow update rates from modulating elements such utilizing spatial light modulators (SLM) which only clock at about 60 Hz, e.g. Ref [39]. Thus, the rationale is to combine Fourier-optics based signal processing with integrated photonics capabilities, which thus enables i) fast update rates of state-of-the-art integrated photonics such as modulators, phase shifters, and photodetectors delivering a foundry-ready capability of 10s of GHz speeds, hence a speed up of about $10^{10}/10^2 = 8$ orders of magnitude, and ii) significantly reduced form-factors of square millimeters (for the active chip at least) rather than 0.1–1 square meter of a free-space 4F systems.

However, the shortcoming to an integrated photonics-based Fourier filter is to trade-off parallelism as further discussed below. Hence, naturally those applications that require a short delay inference and low computational footprint such as network-edge processing benefit most from an integrated FFT-based optical processor [12]. The rationale for this paper is to explore the performance of photonic tensor operations via optical Fourier transforms (Figure 1). We first introduce the general paradigm including several optical and photonic technology options along with two alternative concepts of a parallel and a serial processor, and thence analyze the optical FFT (OFFT) performance with respect to phase and temperature sensitivity before concluding with a size-scaling performance comparison against a GPU.

2 System architecture, design and operation

A unique feature of an electromagnetic wave is its ability to execute the mathematical operation of both addition and

subtraction as it propagates. Such wave interference-based arithmetic forms the basis of several optical effects and system to include, but not limited to, holography, phased array antennas, and interferometric microscopy. Notably, these systems are energy-wise 'passive' or reprogrammable filters, with the only energy consumed by these arithmetic operations in case when losses are incurred of the traversing optical beam or signal. Incidentally, frequency domain filtering is a technique that is deployed since the middle of the last century with Fourier optics. In such an optical processor, a Fourier lens performs an FFT by converting the signal (e.g. an image, RF data, …) into the frequency domain where filtering can occur. The so-filtered and hence processed signal is thence being converted back into the spatial domain with a second lens, i.e. inverse FFT (IFFT). Indeed, executed in free-space operating on the full image such as 1000 × 1000 pixels, namely one million parallel channels, these systems are highly parallel but also bulky, as mentioned above.

Conceptually, the rationale to utilize optical interference to perform an FFT was first introduced by Marhic [40], where originally a star coupler is used to perform addition and subtraction and length differences are used to rotate phase. These concepts are then applied to a convolutional neural networks (CNN) analysis in [31, 41]. However, the phase sensitivity of a star-coupler requires a stable phase and hence real-time phase measurements and active stabilization may be needed adding to complexity, and may reduce the benefits from an optical NN whose core latency benefit is avoidance of OEO conversions. A second optical FFT option is an Mach Zehnder interferometer (MZI) based-mesh network [42, 43], and a third uses the Cooley-Tukey method followed here [44–46]. For these options, their FFT scaling, in order of complexity, are as follows; Star-Coupler ~1; the Cooley-Tukey method ~ ½$N\text{Log}_2N$, and the MZI ~ ½ $N(N\text{-}1)$. This seem to favor the star-coupler, which, however, is rather phase noise sensitive and with increasing N needs to address shrinking design window, and it seems N~20 maybe be a realistic design limit [31].

By using an OFFTs for convolution instead of GPUs, a system can be built to take advantage of the energy efficient arithmetic of wave interference to perform the convolutions of the CNN (Figures 1 and 2). Here the CNN is comprised of FFT point-wise dot product multiplication in the Fourier domain (i.e. frequency filtering), and inverse OFFT. The challenge of this system is that it also requires phase coherence, a sensitivity analysis is discussed below. The OFFT is built on three passive components: the two-by-two coupler is used for addition and subtraction, waveguides with short path differences are used for phase rotation, and waveguides with long path differences are used for signal delay (i.e. spirals) [45]. While in principle an OFFT network could be created with perfect phase alignment at a specific temperature, in practice active phase calibration is required to compensate for fabrication and temperature variance. This phase calibration is normally accomplished with heating elements placed along one of the waveguide paths of each waveguide pair. The Cooley-Tukey FFT requires two operations: addition and multiplication by a phase. The two-by-two optical coupler forms the principal addition equation of the OFFT, Eq. (1).

$$\beta_1 = \frac{1}{\sqrt{2}}(-\alpha_1 + \alpha_2)$$

$$\beta_2 = \frac{1}{\sqrt{2}}(\alpha_1 + \alpha_2) \tag{1}$$

where β_1 and β_2 are the outputs and α_1 and α_2 are the outputs of the 2 × 2 coupler. The phase multiplication can be implemented optically via a difference in phase, Eq. (2).

$$\varepsilon_{xy} = \exp(-i2\pi xy/N) \tag{2}$$

This so-termed butterfly pattern (lower right, Figure. 1a) of the Cooley-Tukey FFT only requires passive optics, ignoring possible required phase control. However, unlike OFDM for which OFFTs have been used before [47], a convolution (generally) operates on spatial data and hence requires a two-dimensional (2D) FFT. This i x j 2D FFT can be realized via row and column operations of i length j FFTs plus j length i FFTs and requires 2 j FFTs of length j each for an exemplary square matrix. In the OFFT there is a choice between implementing a large 2D FFT network directly or implementing a smaller 1D FFT in the complex domain and using it repeatedly in time for each row and column operation [48]. Note, Eqs. (1) and (2) also hold for complex signals. Realizing a complex OFFT, an additional reference signal must be mixed with the output prior to digitization to determine both phase and amplitude, similar to an optical heterodyne Quadrature Phased Shift Keying (QPSK) receiver. Alternatively, the phase can be measured by phase-difference as in an optical Differential Phased Shift Keying (DPSK) receiver, however requiring two cycles for the OFFT including transmitting a known signal for phase determination and then transmitting the actual data.

Knowing both the real and imaginary part of the FFT at the output allows the OFFT to be reused across several cycles to create the 2D FFT required for the CNN (Figure. 2a). Additionally, the 1D Cooley-Tukey algorithm can be divided over multiple cycles due to its recursive nature, which allows scaling down the size of the OFFT properly for the application in use, with a tradeoff between the number of ADCs and DACs and their operating speeds. Indeed, many ADCs and DACs can be replaced by a smaller number

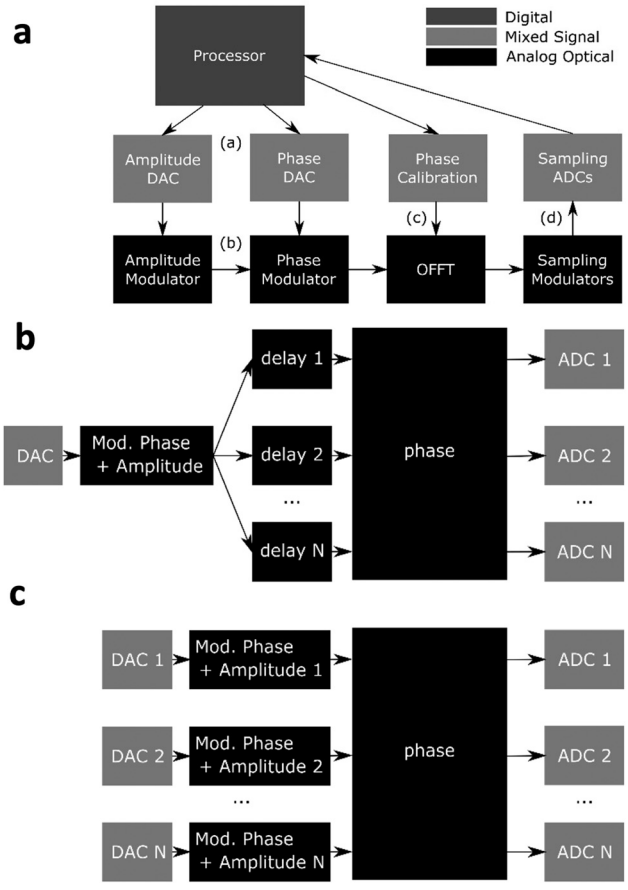

Figure 2: Two options of implementing an optical temporal FFT. Photonic CNN paradigm utilizing Fourier-optics based on integrated photonics. a, Block diagram of the photonic processor utilizing optical FFT to perform convolutions showing the data flow from the processor to the amplitude and phase Digital to Analog Converters (DACs) (i), being modulated onto an optical carrier (ii) flowing through the phase calibrated OFFT network (iii) and being converted back into the digital domain (iv) with sampling Analog to Digital Converters (ADCs) and optional sampling modulators. b, The optical convolution can be run in serial with a single DAC and optical delay in the photonic network, or c, with *N* parallel DACs and no additional delay in the photonic network. Seamless all-photonic DACs could enable reduced OFFT and CNN design complexity, thus enabling higher scaling potential while reducing power consumption due to eliminating the parasitic O-E-O conversion [58].

operating at a higher speed, and considering delay presents another design decision [32]. Overall, the optical FFT tensor processor architecture can be serial, with a single sampling modulator (Figure 2b) and optical delay, or alternatively parallel with N sampling modulators (Figure 2c). Whereas the serial case requires less power, with only one DAC, it also has a lower convolution rate, since only one convolution result can be realized within the period of the longest delay path.

3 Results and discussion

Next we evaluate and discuss the sensitivity of the performance of an examplary $N = 4$ OFFT in silicon photonics numerically (Figure 3a) and extrapolate the performance of a CNN based on a 2D OFFT including performance as a function of N scaling (Figure 4). Interestingly, the FFT delay decreases linearly with increasing modulation rate, thus leading to a more compact system for higher data rates. The on-chip portion of the OFFT consists of cascaded delayed interferometers and passive components such as directional couplers, y branches, straight and spiral waveguides, and operates on time domain signals (Figure 1a). We opt for silicon on insulator (SOI) waveguides as a reconfigurable PIC platform and reliable fabrication process [49–53], but for effects limiting the number of convolutions (N) such as power-scaling, silicon nitride photonics would be better suited. The length of the OFFT MZI phase shifter is determined by the system frequency (i.e. signal modulation), $\boldsymbol{f_s} = 10\ \boldsymbol{GHz}$, then the required time delay becomes $\boldsymbol{T_{Delay}} = \frac{1}{\boldsymbol{f_s}}$ and the physical length, $\boldsymbol{d}_{10} = (\boldsymbol{c}/\boldsymbol{n})\boldsymbol{T_{delay}} = 12\ \boldsymbol{mm}$, hence the MZIs in the waveguide lengths for the delays are trivially computed to be $T/2$ (T/4) = 3 (6) mm in length. The stage-1 MZI must compensate for phase shifts created by fabrication and temperature variances; the relative phase tuning range should cover $\boldsymbol{\pi}/2$ requiring thermal control: with $\frac{\boldsymbol{dn_{eff}}}{\boldsymbol{dT}} \approx \frac{\boldsymbol{dn}}{\boldsymbol{dT}}$ and Silicon's thermo-optical coefficient $\left(\frac{\boldsymbol{dn}}{\boldsymbol{dT}} = 1.9 \times 10^{-4}\boldsymbol{K}^{-1}\right)$, the phase change is given by:

$$\Delta\phi = \frac{\pi}{2} = \frac{2\pi}{\lambda}\frac{dn}{dT}\Delta\, TL \tag{3}$$

For the stage-2 ($N = 4$) MZIs the temperature stability required is 4.2 K, thus, temperature control must not be ignored, and hence we integrate resistive heaters on one of the MZI arms to tune the desired refractive index change. We selected a minimum waveguide bending radius of 50 μm to keep the radiative bending losses low, resulting in a delay-line spiral area of 3.9×10^{-3} mm^2 for T. Sampling is required to obtain the frequency components of the transfer function of the FFT, realized with back-end electro-optic modulators in silicon photonics e.g. Ref [16] or alternatively with emerging modulator-concepts featuring heterogeneous integration of strong-index changes materials such as transparent conductive oxides featuring strong light matter interaction near epsilon near zero (ENZ) operating points [54, 55], and micrometer compact MZI ITO-based modulators [21–23], see also work from the Wong and Brongersma groups (see Figure 4 for nanophotonics-

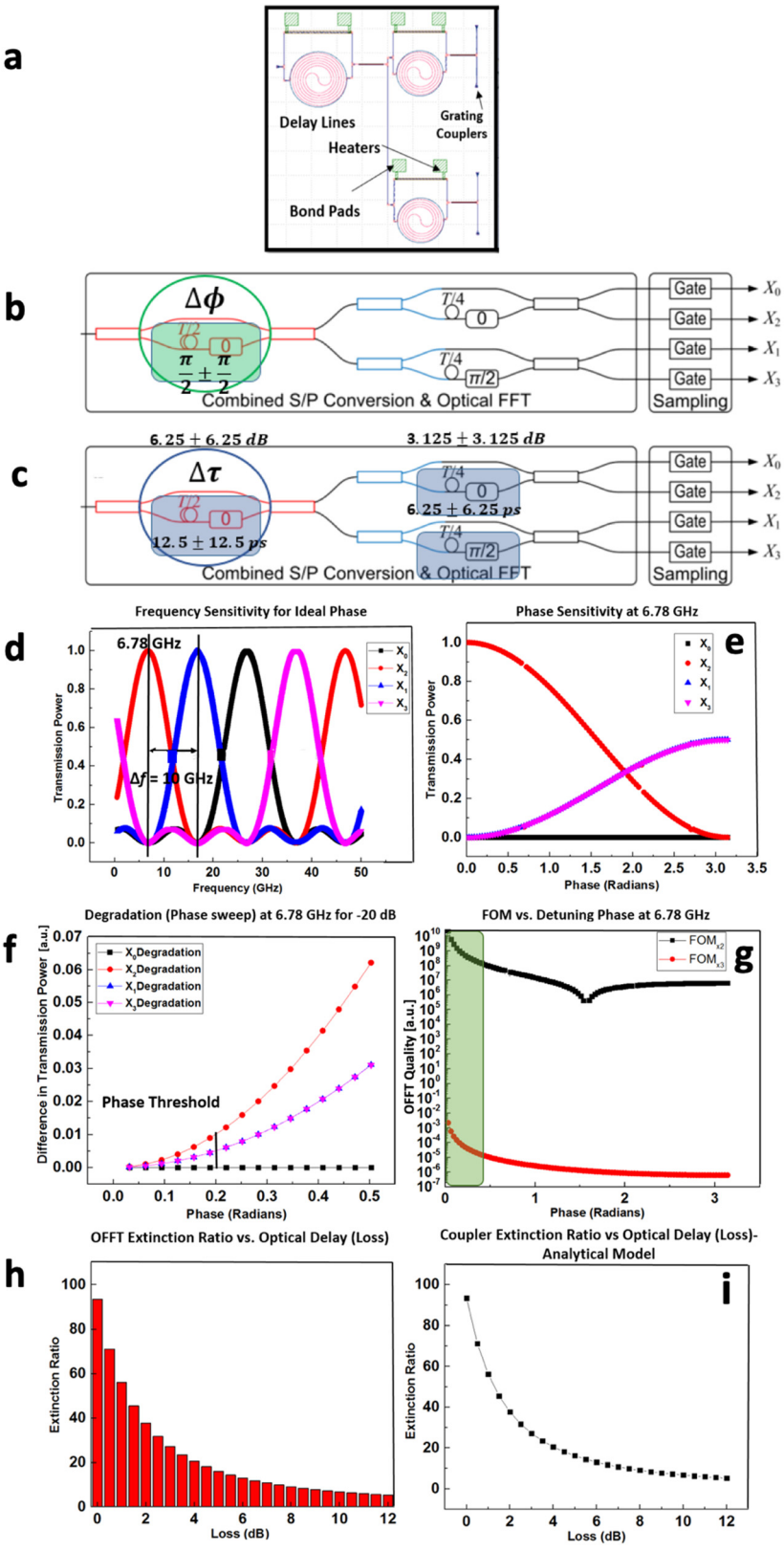

Figure 3: OFFT sensitivity analysis with respect to phase coherence and amplitude loss. a, Layout of an $N = 4$ OFFT design in silicon photonics for numerical analysis (Lumercial: FDTD, MODE, INTERCONNECT); λ_{signal} = 1,550 nm, $n_{WG-mode}$ = 2.5, standard SOI (epi = 220 nm, BOX = 2 ***μm***. The length of the shorter arms of the MZI for the first stage is 500 ***μm*** and for the second stage is 440 ***μm***. b, Phase sensitivity analysis of the OFFT's first stage interferometer. **c,** Time delay sensitivity from T/4 ± T/4 (i. e. 12.5 ± 12.5 ps) corresponding to an optical loss from 6.25 ± 6.25 dB for the first stage MZI. **d,** Frequency sensitivity analysis on the transmission power (transfer function) of OFFT at ideal phase, **e,** Phase sensitivity analysis on the transmission power at probe frequency of 6.78 GHz **f,** Degradation for −20 dB tolerance and **g,** FOM as a function of detuning phase at 6.78 GHz probe frequency- where X_2 has the ideal FOM while X_3 is minimal leakage from X_3 bin (similar to X_0 and X_1, not shown). $X_i = i^{th}$ ouput. Green-shared region denotes −20 dB channel-to-channel separation threshold typical for telecommunication applications [45]. **h,** Extinction ration (ER) of the OFFT as a function of physical optical loss from the spiral waveguides (delay lines), **i,** Analytical exponential fit based on an ideal coupler.

enabled OFFT and convolution performance outlook). Selecting the SOI modulators, for now, results in a total OFFT area to about 0.019 mm^2 (N = 4). The OFFT design allows placing the EOMs either before or after the FFT butterflies. However, in order to minimize detuning phase, delay, and optical loss differences across the four

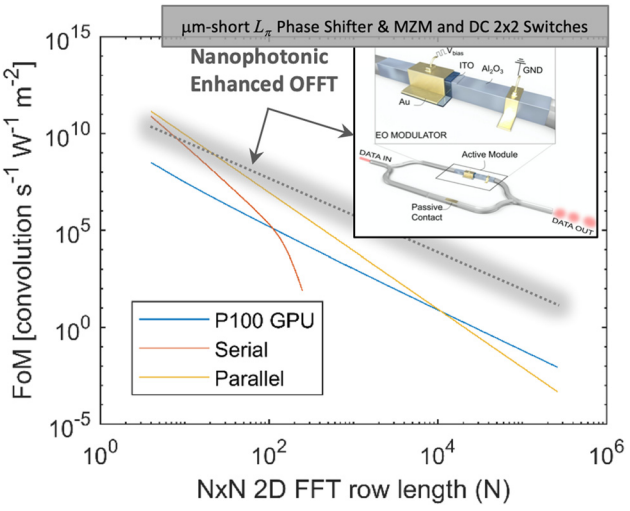

Figure 4: Power-Area-Speed performance analysis of the core of the CNN engine based on OFFT. A model of the figure of merit, Convolutions s^{-1} $W^{-1}m^{-2}$, vs N shows the OFFT architecture outperforming the NVIDIA P100 in for $N < 10^2$ in a serial configuration and less than $N < 10^4$ in parallel configurations. Details are: FLOPS per convolution: $20N^2\log_2(N) + N^2$. ADC(DAC): 56(100) GSa/s @ 2(2.5)W power dissipation; optical component losses: first spiral = 0.7 dB; 2 × 2 directional coupler loss = 1.0 dB; y-splitter and combiners = 3.0dB; grating coupler (freespace-to-chip) = 4.0 dB; electro-optic modulator = 3.5dB. Inset: nanophotonic device enhancements enable micrometer-short optoelectronic devices such as i) π-phase shifters leading to compact MZI modulators [21–23] or electro-absorption modulators [19, 20], and also to <10 μm-short 2 × 2 directional couplers [56] allowing for improved N-scaling of OFFT-based convolutions. GPU processing times obtained by fitting runtime.

modulators should appear at the end of the last stage of the OFFT to ensure synchronous sampling. To avoid power mismatch loss from the difference in waveguide length of the cascaded MZIs, 'wavy' waveguide bends are added to the shorter arm of the interferometers to compensate for the power loss at the output of the couplers and are executed with a Python script feeding into a CAD-tool for photonic waveguide placement.

Next, we analyze the sensitivity of the OFFT transfer function by exploring both phase- and signal delay-noise (Figure 3b, c); conceptually the FFT separates frequency contributions of the temporal input signal. Since the system frequency is examplary assumed to be 10 GHz, the frequency spacing of the OFFT output channels match this closely, but the exact location of the probe frequency for which the maximum transmission is obtained for different outputs is a function of time delay and can vary across the outputs. This will create a variation in every four transmission peaks at a specific frequency the contributions from the neighboring channels (X_i = OFFT output channel number, e.g. i = 1-4 for N = 4, Figure 3D). Analyzing this transfer function and the extinction ratio (ER) of the cascaded interferometers, the discrete nature of the OFFT yields intrinsic quantization errors in frequency and sampling artifacts that are a function of the phase. We analyze the phase- and time delay, and optical loss in particular that of the lower arm of the interferometer in the first stage of the OFFT (Figure 3d, e), and evaluate the change in the transfer functions of the output. This location is particularly impactful on the transfer function, since it has the highest oscillation and narrow spacing in the frequency domain [45]. Our observable is the frequency detuning of the maximum point of the transfer function, which, for instance, for output port X_2 appears near 6.8 GHz (Figure 3d). At the ideal case for vanishing phase detuning, the entire power exits at the second branch X_2 of the top interferometer due to the additional phase, i.e. the relative phase difference in bidirectional coupler, where it is sampled by a modulator.

With phase detuning away from the ideal design point, however, the energy of the output transfer functions can leak to the neighboring ports (Figure 3e); for instance, the power of the first interferometer shifts to the top of the second stage lower interferometer exiting of at the next frequency probe (16.8 GHz), since each output bin in the frequency domain has a 10 GHz spacing, determined by the chosen system frequency. By sweeping the frequency, we can find points at which the OFFT has the highests transmission, defined by time delay. Hence the effective impact of the phase detuning is a lateral frequency shift. In the ideal case of no phase variation and at different probe frequencies (frequency at which each OFFT output has a maximum transmission in power), full transmission can be achieved. However, deviations in phase change the output amplitudes respectively where the maximum transmission decreases until it vanishes, while the next full transmission from another output port becomes dominant, which occurs as a cyclic behavior. Note, that the value of 6.78 GHz is not significant by itself, as it can be shifted by detuning the delay lines. However, to determine the overall quality of this cascaded system of interferometers, we set a threshold for the phase-detuned power ratio (i.e. frequency power leakage) between target-to-neighboring output ports; here we select a −20 dB channel-to-channel separation threshold, which is typical for telecommunication applications (green-shared region (Figure 3g) [45]. Given this threshold, the maximum phase tolerance is < 0.2 radians to ensure acceptable spectral leakage, i.e. channel crosstalk (Figure 3f). Physically this range corresponds to a small (ΔT = 0.54 K) temperature change that the waveguide index must tolerate to keep within the selected attenuation

threshold, consequently demanding phase control. One potential approach is to place the OFFT chip in an ambient chamber with temperature isolation such that heat could be transported to only the specific areas as desired. Alternatively, control loops and temperature stabilization of the chip could also be employed. Indeed, we observe a nonlinear phase error, which is likely due to nature of cascaded interferometers and their phase sensitivity with respect to physical delay lines.

To probe the effects of phase detuning errors and distortions in the signal further, we study the difference in the transmission power ($\boldsymbol{P_{degradation}}$) as a function of phase (Figure 3f,g). Thus, the signal-to-noise ratio $\boldsymbol{SNR} = \frac{(\boldsymbol{P_{out}} - \boldsymbol{P_{degradation}})}{\boldsymbol{P_{degradation}}}$ is obtained by taking the difference in the transmission output power values relative to the ideal zero phase, and, thus, determines the performance quality of the OFFT in response to optical phase noise. To help guide the discussion and analyze this sensitivity qualitatively, we define the power mismatch ratio and FFT quality as follows:

$$P_{mismatchRatio} = \frac{P_{out1}}{P_{out2}}(\phi) \tag{4}$$

$$OFFT - Quality = \frac{SNR}{P_{mismatchRatio}} \tag{5}$$

The rationale for these definitions is that the smaller the power mismatch ratio (deviation from unity), and the higher the SNR in the system, the higher the quality of the OFFT as a function of detuning phase. As a result, the system performs best at the ideal zero phase case as expected for X_2 since the power mismatch ratio between X_2 and X_0 is close to 1, where Eq. (5) is maximized. For the case of X_1 and X_3 however, their SNR values are low, despite the power mismatch ratio being close to unity 1, since their transmission is minimal for frequency contribution at 6.78 GHz, and naturally the maximum OFFT quality for all four outputs aligns with the design probe frequency value for each frequency bin. As the phase is detuned the OFFT quality degrades for X_2 (Figure 3g).

For ideal OFFT filtering and optimal transmission, the output power of the cascaded MZI arms must match. This, however, is challenging since in the OFFT design the MZI arms have different physical lengths, hence, in order to understand how the difference in length changes the quality of the OFFT output, the MZI extinction ratio (*ER*) was analyzed by sweeping the delay and additional loss in the first and correspondingly second stage of OFFT (Figure 3h,i). The extrinction ratio is defined as $\boldsymbol{ER} = \frac{\boldsymbol{P_{max}}}{\boldsymbol{P_{min}}}(\boldsymbol{\gamma_{loss}})$where $P_{\max}$ ($P_{\min}$) is the maximum (minimum) power at the output of the OFFT, and the delay lines corresponding to loss values were altered across the lower arm of the cascaded MZIs. Note, that in the second stage the delay is half of the first stage, and so is the loss. This is important for consistency and symmetric operation in the overall system design. As proven analytically by the ideal coupler's ER behavior in the lower arm (Figure 3i), the loss increases exponential with waveguide length as expected. However, the loss increases if the length imbalance increases, due to the extra power mismatch between the MZI arms, impacting the quality of the OFFT's transfer function. Thus, the aim is to maximize ER, similar to modulators and 2×2 direction coupler electro-optic switches e.g. a < 10 μm compact option is Ref [56], but with the difference of improving the power mismatch between the MZI arms rather signal device on-off ratio. Another significant scaling factor in the OFFT resides in increasing the optical losses from a growing the number of branches and waveguide delay lengths at high *N*. The optical power at each output must be sufficiently high to meet the noise requirements for the number of bits at the detector. With the total loss for a single output being the sum of the losses in dB at each $\log_2\boldsymbol{N}$ stage, in the serial architecture, the longest arm of a stage passes through a 3-dB branch and an MZI-spiral with a length that scales with $\boldsymbol{N}$. At high *N* the loss becomes dominated by the $\boldsymbol{N}\log_2\boldsymbol{N}$ scaling of the spirals.

Indeed, any analog processor, photonic versions included, is challenged by obtainable SNR. Fundamental options to raise the SNR are to increase the available optical power front-end, in case losses are predominant. Alternatively, optical engines for NN's such as CNNs could be used iteratively, processing one layer at a time. This enables de-coupling the SNR at each layer from the final output result by, for instance, electronic signal restoration, with a possible drawback introduced being latency. Nonetheless, noise is not only a limiting factor, but bears also an opportunity for NN usages; for instance, training [57] an NN with noise and performing inference tasks while dropping the absolute accuracy by about 2–3% also makes the system more robust against physical noise since the NN was conditioned with noise 'stress' [34]. Incidentally, the small-kernel algorithms such as the Winograd transformation offer an interesting alternative to the FFT filtering approach, given that many CNNs are optimized for small kernel ($<13 \times 13$) sizes.

4 Performance: sample scaling, power and speed

The OFFT in silicon photonics becomes a network of delay waveguides, Y-branches, MZIs, and heater-calibrated

phase delay waveguide segments (Figure 3a–c). In the serial case (Figure 2a), delay is implemented with spiral delay stages. The spirals scale in area proportional to their length. The length of the first spiral is the greatest and they diminish in length with $(1/2)^k$, where k is the stage index. There are $\log_2(N)$delay stages each with 2^k spirals. Even though the number of spirals doubles with each stage, the area stays constant due to the spiral length halving with each stage. Thus, the area relative to the first spiral scales with $\log_2(N)$ and the first spiral with scales with N for a total area scaling of $N\log_2(N)$.

It is apparent that due to the passive nature of the OFFT network, the primary power consumer in a small-N OFFT is found in the conversion between the digital and analog domains. If the OFFT were directly connected to an analog fully connected NN, or some other analog processor, the only power consumed by the OFFT would come from optical propagation losses, phase compensation, and the coherent optical source. However, todays dominant computer architecture is digital, and to be practical, the OFFT implementation must interface to the digital domain. The power consumption analysis becomes an analysis of digital and analog conversions. Recent developments in DAC and ADC designs now also allow leaving the signal in the optical domain altogether (no OEO conversion required), thus reducing system design complexity [58].

The FFT data capacity, the number of bits that can be propagated through the system, depends on the modulation type; assuming QAM 256 for a high SNR channel with a bandwidth of 10 GHz the upper bound for bandwidth is 80 Gbps for a single OFFT channel and 320 Gbps for $N = 4$. While we have analyzed the sensitivity and performance for $N = 4$, it is interesting to ask how larger systems scale. Increasing the number of FFT-samples (N), the OFFT grows with $(N - 1)$ cascaded delayed interferometers and $2(N - 1)$ couplers. Unlike an electronic FFT however, which scales with approximately$5N\log_2N$, the optical FFT will need to compensate for increasing optical losses with greater optical power. Our FFT scaling analysis shows performance peaks for the OFFT for small N, which outperforms an electronic (NVIDIA P100 GPU) for $N < 200$ (Figure 4) [59]. The results of the performance-scaling characteristics of the OFFT using the highest speed DAC and ADC found in literature today, and comparing against NVIDIA P100 GPU, shows up to 20 times improved performance even with the high power consumed by the converters. The NIVIDIA P100 performs 1.6 TFLOPS during single precision 1024 length FFT [59]. Assuming the 1D FFT requires $2N\log_2(N)$ multiplication operations and $3N\log_2(N)$ addition operations there will be a total of $5N\log_2(N)$ FLOPS per 1D FFT of length N. To generate a 2D FFT with an edge length of N from a 1D FFT there will be $2N$ 1D FFTs of length N for each 2D FFT. To complete the convolution there will be one N^2 multiply in the frequency domain and one inverse 2D FFT to return to the spatial domain. Then the number of FLOPS to convolutions becomes $20N^2\log_2(N) + N^2$. For the exemplary P100 this results in a convolution rate of 7 KHz with $N = 1024$ and 150 KHz with $N = 256$. Modern DACs operate at about 100 GSa/s and consume 2.5 W [60], while ADCs show 56 GSa/s consuming 2 W per channel [61]. Assuming integrated germanium photodetectors [62] with a reverse bias of 8 V and 250 μW of optical power, the power consumption in each photodiode is approximated as 2.4 μW. Using these assumptions, we model both serial and parallel implementations of the OFFT convolutional architecture and compared them to the NVIDIA P100 GPU using a 2D OFFT convolutional filter (1-stage of a CNN) system performance FoM defined as number of convolutions $s^{-1}\,W^{-1}\,m^{-2}$ (Figure 4). The results of the analysis show that with a small convolution size the photonic approach is outperforms electronics (here the P100 GPU). However, the advantage diminishes as the convolution scales, which is due to the approximate scaling of the FoM with $1/(N^4\log_2N10^{(N\log_2N)/10})$ in the serial photonic case, $1/(N^4\log_2N)$ in the parallel photonic case, and $1/(N^2\log_2N)$ in the electronic case. Exemplary, the contribution to the serial photonic FFT scaling are from $1/(N\log_2N)$ in area,$1/(10^{(N\log_2N)/10})$ in optical power, $1/N$ in ADC power, $1/N^2$ in samples to area. Shortly beyond $N = 10^2$ the power efficiency of the P100 overtakes the serial photonic approach and near $N = 10^4$ the P100 passes the parallel OFFT architecture. This model assumes a constant TFLOPS performance for the P100 independent of N. A more realistic model would reduce the TFLOPS performance of the GPU at higher N due to pipelining and memory bandwidth limits. Note, the results of Figure 4 are considered an upper bound and assume noise stability of the OFFT, as discussed in Figure 3, can be ensured. While these are encouraging initial results, further improvements are anticipated as optoelectronic components evolve including nanophotonics, plasmonics or metamaterial-based principles to deliver <1fJ/bit level devices [63]. If such level of efficiency was achievable, the photodiode and electronic ADC in our previous analysis could potentially be replaced with a set of 256 single bit receivers operating at 10 s of femtojoules per bit. In this case, the power at each 8 bit 56 GSa/s receiver could be reduced to as little as 5 mW.

Aside from capacity-density scaling, there are several system-level design optimization that should be addressed for improved system performance. Detailed discussions go

beyond the scope of this work, but there are two critical factors that are becoming the next bottlenecks; i) optical analog processors capable of TOPS/s yet approaching POPS/s throughput are, however, challenged with providing sufficiently fast data input/output (I/O). While the large-array detector read-out is not a factor for integrated photonic system such as considered here, the streaming capability even of top-of-the-line FPGAs of <10Tbps is approaching demands for front-end I/O of optical accelerators. ii) domain crossing from digital to analog require additional converters (DAC/ADC). iii) latency and energy to deliver rapidly changing kernel weights from the memory may become prohibitively cumbersome, and local photonic memory solutions such as provided by phase-change materials must be considered in a system as well.

5 Conclusion

In conclusion, we have introduced a generalized photonic tensor-operation processor harnessing the natural parallelism of optics to performing the overhead-heavy convolution operation as (simpler) dot-product multiplications in the Fourier domain. Domain transformations into- and out-of the Fourier domain can be performed passively and optically, here via a 'Butterfly' FFT (Cooley-Turkey Method). We obtain the transfer function of this photonic FFT by monitoring the frequency bins at the output ports of the FFT sampled by electro-optic modulators. A sensitivity analysis shows that the thermal operating range is achievable, yet somewhat narrow, in order to adhere to telecommunication-relevant phase thresholds. Unlike electronics, here the number of FFTs being processed per second only depends on the time-of-flight of a photon through the millimeter short photonic chip, and is hence in the *~ps* range, which can be considered real-time for compute systems. As such, we find the performance (number of FFTs or convolutions performed data per second, power, and areal footprint) outperforms state-of-the-art graphical processing units GPUs) by 2 orders of magnitude for small size input data. Nanophotonics component level developments of modulators and directional coupler switches, for instance, can enable even higher performance enabled by a more efficient way of engineering strong light-matter-interactions, and can also used to engineer the nonlinear activation function the CNN. Taken together such optical FFT-based tensor processor shows how integrated photonics enables data processing by simple routing light through a configured network (in-the-network-computing), thus enabling photonic domain-specific engines such as for machine intelligence rather than using photons only for interconnect applications.

Acknowledgments: This project was funded by the research and development office (RDO) at the Ministry of Education, Kingdom of Saudi Arabia. Grant no (HIQI-36-2019). The authors also, acknowledge with thanks research and development office (RDO-KAU) at King Abdulaziz University for technical support. The authors acknowledge fruitful discussion with Jonathan George, Hani Nejadriahi, and Mario Miscuglio.

Author contribution: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.

Research funding: Research and Development Office (RDO) at the Ministry of Education, Kingdom of Saudi Arabia (grant no. HIQI-36-201).

Conflict of interest statement: The authors declare no conflicts of interest regarding this article.

References

[1] R. K. Cavin, P. Lugli, and V. V. Zhirnov, "Science and engineering beyond Moore's law," *Proc. IEEE*, vol. 100, no. Special Centennial Issue, pp. 1720–1749, 2012.

[2] N. Jouppi, C. Young, N. Patil, and D. Patterson, "Motivation for and evaluation of the first tensor processing unit," *IEEE Micro*, vol. 38, no. 3, pp. 10–19, May 2018.

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

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural infor- mation processing systems*, pp. 1097–1105, 2012.

[5] S. B. Furber, F. Galluppi, S. Temple, and L. A. Plana, "The spinnaker project," *Proc. IEEE*, vol. 102, no. 5, pp. 652–665, 2014.

[6] S. K. Essera, P. A. Merollaa, J. V. Arthura, et al., "Convolutional networks for fast energy- efficient neuromorphic computing," *Proc. Nat. Acad. Sci. USA*, vol. 113, no. 41, pp. 11 441–11 446, 2016.

[7] M. Miscuglio, G.C. Adam, D. Kuzum, V.J. Sorger "Roadmap on material-function mapping for photonic-electronic hybrid neural networks, " *APL Materials*, vol.7, p. 100903, 2019.

[8] H. J. Caulfield, J. Kinser, and S. K. Rogers, "Optical neural networks," *Proc. IEEE*, vol. 77, no. 10, pp. 1573–1583, 1989.

[9] J. W. Goodman, *Introduction to Fourier Optics*, Roberts & Company, 2005.

[10] X. Lin, Y. Rivenson, N. T. Yardimci, et al., "All-optical machine learning using diffractive deep neural networks," *Science*, vol. 361, pp. 1004–1008, 2018.

[11] 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, 2019.

[12] V. J. Sorger, "Photonic Convolutional Processor of for Network Edge Computing," ONR Electronic Warfare, (accessed Jan-2020).
[13] P. Prucnal, and B. Shastri, *CRC 'Neuromorphic Photonics'* 2018.
[14] A. N. Tait, A. X. Wu, T. F. de Lima, et al., "Microring weight banks," *IEEE J. Sel. Top. Quantum Elect.*, vol. 22, no.6, pp. 312–325, 2016.
[15] A. N. Tait, T. F. De Lima, E. Zhou, et al., "Neuromorphic photonic networks using silicon photonic weight banks," *Sci. Rep.*, vol. 7, no. 1, pp. 7430, 2017.
[16] Y. Shen, N. C. Harris, S. Skirlo, et al., "Deep learning with coherent nanophotonic circuits," *Nat. Photon.*, vol. 11, no. 7, p. 441, 2017.
[17] D. Brunner, M. C. Soriano, C. R. Mirasso, and I. Fischer, "Parallel photonic information processing at gigabyte per second data rates using transient states," *Nat. Commun.*, vol. 4, p. 1364, 2013.
[18] A. N. Tait, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, "Broadcast and weight: an integrated network for scalable photonic spike processing," *J. Lightwave Technol.*, vol. 32, no. 21, pp. 3427–3439, 2014.
[19] V. J. Sorger, D. Kimura, R.-M. Ma, X. Zhang, "Ultra-compact silicon nanophotonic modulator with broadband response," *Nanophotonics*, vol. 1, no.1, pp. 17–22, 2012.
[20] M. H. Tahersima, Z. Ma, Y. Gui, et al., "Coupling-enhanced dual ITO layer electro-absorption modulator in silicon photonics," *Nanophotonics*, vol. 8, p. 9, 2019.
[21] R. Amin, R. Maiti, C. Carfano, et al., "0.52 V-mm ITO-based mach-zehnder modulator in silicon photonics," *APL Photonics*, vol. 3, p. 12, 2018.
[22] R. Amin, R. Maiti, J. K. George, et al.,"A lateral MOS-capacitor enabled ITO Mach- zehnder modulator for beam steering," *J. Lightwave Technol.*, 2019, https://doi.org/10.1109/jlt.2019.2956719.
[23] R. Amin, R. Maiti, Y. Gui, et al., "Broadband sub-λ GHz ITO plasmonic mach-zehnder modulator on silicon photonics," *Optica*, vol. 7, p. 3, 2020.
[24] M. Miscuglio, J. Meng, O. Yesiliurt, et al., *Artificial Synapse with Mnemonic Functionality using GSST-based Photonic Integrated Memory"arXiv preprint*: 02221. 1912.
[25] M. Miscuglio, A. Mehrabian, Z. Hu, et al., "All-optical nonlinear activation function for photonic neural networks," *Opt. Mat. Expr.*, vol. 8, no. (12), pp. 3851–3863, 2018.
[26] K. Liu, S. Sun, A. Majumdar, V. J. Sorger, "Fundamental scaling laws in nanophotonics," *Nat. Scientific Rep.*, vol, 6, p. 37419, 2016.
[27] R. Amin, C. Suer, Z. Ma, et al., "A deterministic guide for material and mode dependence of on-chip electro-optic modulator performance," *Solid-State Electronics*, vol. 136, pp. 92–101, 2017.
[28] R. Amin, C. Suer, Z. Ma, J. Khurgin, R. Agarwal, V. J. Sorger, "Active Material, Optical Mode and Cavity Impact on electro-optic Modulation Performance" Nanophotonics, vol. 7, no. 2, pp. 455–472, 2017.
[29] J. K. George, A. Mehrabian, R. Armin, et al., "Noise and nonlinearity of electro-optic activation functions in neuromorphic compute systems," *Optics Express*, vol. 27, p. 4, 2019.
[30] R. Amin, J. George, S. Sun, et al., "ITO-based electro-absorption modulator for photonic neural activation function," *APL Materials*, vol.7, Atl no. 081112, 2019, https://doi.org/10.1063/1.5109039.
[31] J. R.Ong, C. C. Ooi, T. Y. L. Ang, S. T. Lim, C. E. Png, "Photonic convolutional neural networks using integrated diffractive optics, " *IEEE STQE*, 2020, https://doi.org/10.1109/jstqe.2020.2982990.
[32] H. Bagherian, S. Skirlo, Y. Shen, H. Meng, V. Ceperic, and M. Soljacic, "*On-chip optical convolutional neural networks*," arXiv preprint arXiv:1808.03303, 2018.
[33] Viraj Bangari, Bicky A Marquez, Heidi Miller, et al., "Digital electronics and analog photonics for convolutional neural networks (DEAP-CNNs)," *IEEE JSTQE* 26, 1, pp. 1–13, 2019.
[34] A. Mehrabian, M. Miscuglio, Y. Alkabani, V. J. Sorger, T. El-Ghazawi "A winograd-based integrated photonics accelerator for convolutional neural networks," *IEEE J. of Selected Topics in Quantum Electronics*, vol. 26, no. (1), pp. 1–12, 2019.
[35] S. Colburn, Y. Chu, E. Shlizerman, A. Majumdar, "An Optical Frontend for a Convolutional Neural Network" Applied Optics," Vol. 58, no. 12, pp. 3179–3186, 2019.
[36] M. E. Marhic, "Discrete fourier transforms by single-mode star networks," *Opt. Lett*, vol. 12, pp. 63–65, 1987.
[37] J. J. López, et al., "Planar-lens enabled beam steering for chip-scale LIDAR," *Conference on Lasers and Electro-Optics (CLEO)*, San Jose, CA, pp. 1–2, 2018, https://doi.org/10.1364/cleo_si.2018.sm3i.1.
[38] Z. Hu, M. Miscuglio, J. George, Y. Alkabani, T. Gazhawi, and V. Sorger, "Highly-parallel optical fourier intensity convolution filter for image classification," *FIO, paper JW4A*, vol. 101, 2019, https://doi.org/10.1364/fio.2019.jw4a.101.
[39] e.g. Pluto Holoeye, Available from: (https://holoeye.com/spatial-light-modulators/, online (accessed Jan-2020).
[40] M. E. Marhic,"Discrete fourier transforms by single-mode star networks," *Opt. Lett*, vol. 12, pp. 63–65, 1987.
[41] Z. Wang, K. S. Kravtsov, Y. Huang, P. R. Prucnal, "Optical FFT/IFFT circuit realization using arrayed waveguide gratings and the applications in all-optical OFDM system," *Optics Express*, vol. 19, no. 5, pp. 4501, 2011.
[42] M. Reck, A. Zeilinger, H. J. Bernstein, and P. Bertani, "Experimental realization of any discrete unitary operator," *Phys. Rev. Lett.*, vol. 73, no. 1, p. 58, 1994.
[43] W.R. Clements, P.C. Humphreys,B.J. Metcalf,W.S. Kolthammer,and I. A. Walmsley, "Optimal design for universal multiport interferometers," *Optica*, vol. 3, no. 12, pp. 1460–1465, 2016.
[44] R. Barak and Y. Ben-Aryeh, "Quantum fast fourier transform and quantum computation by linear optics," *JOSA B*, vol. 24, no. 2, pp. 231–240, 2007.
[45] Hillerkuss, D., M. Winter, M. Teschke, et al., "Simple all-optical FFT scheme enabling Tbit/s real-time signal processing," *Opt. Express*, vol. 18, pp. 9324–9340, 2010.
[46] Jiaqi Gu, Zheng Zhao, Chenghao Feng, Mingjie Liu, Ray T. Chen, David Z. Pan, "Towards area-efficient optical neural networks: an FFT-based architecture, " *IEEE/ACM Asian and South Pacific Design Automation Conference* (ASPDAC), Beijing, China, Jan. 13-16, 2020.
[47] D. Hillerkuss, R. Schmogrow, T. Schellinger, et al., "26 tbit s-1 line-rate super-channel transmission utilizing all-optical fast

fourier transform processing," Nature Photonics, vol. 55, no. 6, pp. 364–371, 2011.
[48] R. Gray and J. Goodman, *Fourier Transforms an Introduction for Engineers*, Springer Science & Business Media, 1995, pp. 74–78.
[49] N. C. Harris, J. Carolan, D. Bunandar, et al., "Linear programmable nanophotonic processors," *Optica*, vol. 5, no. 12, pp. 1623–1631, Dec 2018.
[50] A. H. Atabaki, S. Moazeni, F. Pavanello, et al., "Inte- grating photonics with silicon nanoelectronics for the next generation of systems on a chip," *Nature*, vol. 556, no. 7701, p. 349, 2018.
[51] J. Sun, E. Timurdogan, A. Yaacobi, et al., "Large-scale silicon photonic circuits for optical phased arrays," *IEEE J. Sel. Topics Quantum Electron.*, vol. 20, no. 4, pp. 264–278, 2013.
[52] S. Han, T. J. Seok, N. Quack, B.-W. Yoo, and M. C. Wu, "Large-scale silicon photonic switches with movable directional couplers," *Optica*, vol. 2, no. 4, pp. 370–375, 2015.
[53] D. Pe´rez, I. Gasulla, L. Crudgington, et al., "Multipurpose silicon photonics signal processor core," *Nat. Commun.*, vol. 8, no. 1, p. 636, 2017.
[54] Y. Gui, M. Miscuglio, Z. Ma, M. T. Tahersima, V. J. Sorger "Towards integrated metatronics: a holistic approach on precise optical and electrical properties of Indium Tin Oxide," *Nature Scientific Reports*, vol. 9, no. 1, pp. 1–10, 2019.
[55] N. Kinsey, J. Khurgin "Nonlinear epsilon-near-zero materials explained," *Opt. Mat. Exp.* Vol. 9, no. 7, pp. 2793–2796, 2019.
[56] C. Ye, K. Liu, R. Soref, V. J. Sorger, "A compact plasmonic MOS-based 2x2 Switch"*Nanophotonics*, vol. 4, no. 1, pp. 261–268, 2015.
[57] T. W. Hughes, M. Minkov, Y. Shi, S. Fan, "Training of photonic neural networks through in situ backpropagation and gradient measurement," *Optica*, vol. 5, no. 7 pp. 864–871, 2018.
[58] J. Meng, M. Miscuglio, J. K. George, et al., "Electronic bottleneck suppression in next-generation networks with integrated photonic digital-to-analog converters" nature communications," under review, 2019. arXiv preprint: 1911:02511.
[59] NVIDIA. NVIDIA TESLA P100 GPU accelerator [online]. Available: https://images.nvidia.com/content/tesla/pdf/nvidia-tesla-p100-PCIe-datasheet.pdf. (accessed Jan-2020).
[60] H. Huang, J. Heilmeyer, M. Grözing, et al., "An 8-bit 100-GS/s distributed DAC in 28-nm CMOS for optical communications," *IEEE Trans. Microw. Theory Tech*, vol. 63, no. 4, pp. 1211–1218, 2015.
[61] Fujitsu. 56GSa/s 8-bit analog-to-digital converter [Online]. Available: 2010, https://www.fujitsu.com/downloads/MICRO/fma/pdf/56G_ADC_FactSheet.pdf.
[62] Y. Zhang, S. Yang, Y. Yang, et al., "A high-responsivity photodetector absent metal-germanium direct contact," *Optics Express*, vol. 22, no. 9, pp. 11367, 2014.
[63] D. A. B. Miller, "Attojoule optoelectronics for low-energy information processing and communications," *J. Lightwave Tech.*, vol. 35, no. 3, pp. 346–396, 2017.