optica

Massively parallel amplitude-only Fourier neural network

Mario Miscuglio,[1] Zibo Hu,[1] Shurui Li,[2] Jonathan K. George,[1] Roberto Capanna,[3] Hamed Dalir,[4] Philippe M. Bardet,[3] Puneet Gupta,[2] and Volker J. Sorger[1,*]

[1]Department of Electrical and Computer Engineering, George Washington University, Washington, DC 20052, USA
[2]Department of Electrical and Computer Engineering, University of California, Los Angeles, California 90095, USA
[3]Department of Mechanical and Aerospace Engineering, George Washington University, Washington, DC 20052, USA
[4]Optelligence LLC, Alexandria, Virginia 22302, USA
*Corresponding author: sorger@gwu.edu

Received 27 August 2020; revised 3 November 2020; accepted 16 November 2020 (Doc. ID 408659); published 18 December 2020

Machine intelligence has become a driving factor in modern society. However, its demand outpaces the underlying electronic technology due to limitations given by fundamental physics, such as capacitive charging of wires, but also by system architecture of storing and handling data, both driving recent trends toward processor heterogeneity. Task-specific accelerators based on free-space optics bear fundamental homomorphism for massively parallel and real-time information processing given the wave nature of light. However, initial results are frustrated by data handling challenges and slow optical programmability. Here we introduce a novel amplitude-only Fourier-optical processor paradigm capable of processing large-scale ∼(1000 × 1000) matrices in a single time step and 100 μs-short latency. Conceptually, the information flow direction is orthogonal to the two-dimensional programmable network, which leverages 10^6 parallel channels of display technology, and enables a prototype demonstration performing convolutions as pixelwise multiplications in the Fourier domain reaching peta operations per second throughputs. The required real-to-Fourier domain transformations are performed passively by optical lenses at zero-static power. We exemplary realize a convolutional neural network (CNN) performing classification tasks on 2 megapixel large matrices at 10 kHz rates, which latency-outperforms current graphic processing unit and phase-based display technology by 1 and 2 orders of magnitude, respectively. Training this optical convolutional layer on image classification tasks and utilizing it in a hybrid optical-electronic CNN, shows classification accuracy of 98% (Modified National Institute of Standards and Technology) and 54% (CIFAR-10). Interestingly, the amplitude-only CNN is inherently robust against coherence noise in contrast to phase-based paradigms and features a delay over 2 orders of magnitude lower than liquid-crystal-based systems. Such an amplitude-only massively parallel optical compute paradigm shows that the lack of phase information can be accounted for via training, thus opening opportunities for high-throughput accelerator technology for machine intelligence with applications in network-edge processing, in data centers, or in pre-processing information or filtering toward near-real-time decision making.

https://doi.org/10.1364/OPTICA.408659

1. INTRODUCTION

In the recent years, deep learning has thrived due to its ability to learn patterns within data and perform intelligent decisions, superior in some cases to human [1–3]. Convolution neural networks (CNNs) lie at the heart of many emerging machine learning applications, especially those related to the analysis of visual imagery. From a neural network (NN) point of view, a CNN extracts specific features of interest, using linear mathematical operations—convolutions—which combine two pieces of information, namely feature map and kernel, to form a third function (transformed feature map). Interestingly, these convolution layers are responsible for consuming the majority (∼ 80%) of the compute resources during inference tasks [4]. In fact, the convolution between a feature map ($n \times n$) and a kernel ($k \times k$) requires a computational complexity of $O(n^2k^2)$ in the real spatial domain, hence without performing any transformation. This results in a significant latency and computational power consumption, especially for datasets comprising appreciably large feature maps, or requiring deep CNNs for achieving high accuracy [5], even when the network has been trained and the memory initialized. For this purpose, data-parallel specialized architectures such as graphic processing units (GPUs) and tensor processing units (TPUs), providing a high degree of programmability, deliver dramatic performance gains compared to general-propose processors.

When used to implement deep NN, performing inference on large two-dimensional datasets such as images, TPUs and GPUs are rather power hungry and require a long computation time (>tens of milliseconds), which is a function of the complexity of the task and accuracy required, which translates into manifold operations with complex kernel and larger feature map.

As it stands, improving computational efficiency of CNNs is still a challenge, due to the widespread relevance to many applications. Therefore, it is necessary to reinvent the way current computing platforms operate, replacing sequential and temporized operations, and related continuous access to memory with massively parallelized yet distributed dynamical units, pushing toward efficient post-CMOS compute paradigms and system implementations. The intrinsic parallelism, the arbitrary large space–bandwidth product [6] and simultaneous low-energy consumption make free-space optics a particularly attractive candidate for deep learning, computing, and particularly for image classification and pattern recognition using CNNs in real time (low latency). In this context, as late as the 1960s [7], optical filtering and correlations, relying on spatial Fourier transform of images in the frequency domain, were used to extrapolate similarity (specific features) between images and signatures [8]. Subsequently, research groups built optical correlators, convolution processors [9,10] and matrix multipliers [11], with competitive performance for that period, although the tremendous advancement of digital electronics frustrated these efforts. However, early successes of such optical processors did not step beyond prototype stages due to the lack of practical devices for simulation of neural planes [12] and the inability to feed these potentially high-throughput (∼ POPS/s) processors sufficiently with data front end.

Increased data volume and parallel computation requirements together with recent advances in digital display technology poise new opportunities for massively parallel optical accelerators. Optical free-space systems offer processing large matrices (several megapixels), and the CNN-required convolutions can be performed as simpler pointwise multiplications in the Fourier domain where domain crossings (from real to Fourier space, and inverse) are performed passively in a Fourier optics $4f$ system. However, the high parallelism and inherent operations provided by the nature of optical signal is confronted by the rigidity of the current optical tools which lack high-speed programmability. For instance, recent optical systems, used as convolutional layer performing inference after being trained, rely on fixed kernels, realized as 3D printer manufactured diffractive masks [13], or slowly varying (tens of hertz) spatial light modulators (SLMs) [14–16]. On the other hand, state-of-the-art high-speed (gigahertz) programmable metasurfaces and tunable optical phased arrays are still limited in terms of matrix resolution and phase contrast [17,18].

Here, we introduce and experimentally demonstrate a novel compute paradigm based on amplitude-only (AO) electro-optical convolutions between large matrices or images using kilohertz-fast reprogrammable high-resolution digital micromirror devices (DMDs), based on two stages of Fourier transforms (FTs), without the support of any interferometric scheme. Low-power laser light is actively patterned by electronically configured DMDs in both the object and Fourier plane of a $4f$ system, encoding information only in the amplitude of the wavefront. By individually controlling the 2 million programmable micromirrors, with a resolution depth of 8 bit and a speed of 1031 Hz (∼20 kHz with 1 bit resolution), it is possible to achieve reprogrammable operations for (near) real time, which is about 100× lower system latency with respect to current optical convolution accelerators (SLM − based systems[10]) image processing, with a maximum throughput of 4-peta operations per second at 8 bit resolution, emulating on the same platform multiple convolutional layers of a NN.

Additionally, while this study does not dispute the scientific understanding that phase information is more important than the amplitude's in image processing [18], such as in the transmission of a continuous tone picture for preserving its visual intelligibility, for example [19], this study shows that adding robustness to a system via a training paradigm is capable of accounting for lack of information (here phase). That is, leveraging on the robustness of the NN, achieved through hardware-specific training, we show that it is possible to overcome the loss of information related to phase of the modulated radiation, which enables performing intelligent classification in an opportunely trained CNN and concurrently achieving high accuracy [Modified National Institute of Standards and Technology (MNIST) and CIFAR-10 classification] and throughput (10,000 conv/s of ∼2000×1000 large matrices). This architecture experimentally validates the power of an AO $4f$ system optical computing paradigm and further opens up the NN architectures with components that are readably available for parallelly performing intelligent tasks in near real time, such as in free-space communication [20] in data centers for processing data locally at the edge of the network, without communicating across long routes to data centers or clouds.

2. RESULTS

The system architecture typology for realizing the amplitude-only Fourier filter (AO-FF) layer for performing filtering is synergistically realized in optics [21]; a coherent optical image processor is based on a $4f$ system, in which there are four lens focal distances f separating the object from the image plane, intercalated by two Fourier transforming lenses [Fig. 1(a)]. This setup is composed of an input (object) plane, the processing (Fourier) plane, and the output (image) plane. The to-be-processed data and the kernel, which filters them in the Fourier plane, are spatially modulated according an electro-optic transduction. Conceptually, such a free-space approach enables three-dimensional parallelism, which is elegant, since it decouples in-plane (x, y directions) programmability (here provided by the DMD), from the direction of the information flow (z direction).

With the presumption that phase information is more relevant than amplitude information [22], other $4f$ implementations rely on phase modulation based on SLMs[10]. SLMs exploit pixel-wise phase retardation introduced by the variation of the effective refractive index through orientation of birefringent liquid crystals to which a voltage is applied. On the contrary, for our implementation, this transduction is achieved through a DMD, belonging to the family of micro-opto-electro-mechanical system (MOEMS). They consist of micromirror arrays which impose a spatially varying light intensity modulation by rapidly tilting individual micromirrors, which deflect input light. In detail, each pixel of a DMD is comprised of a tilting mirror and a memory unit storing the pattern to be reproduced; the mirror flips according to the digital value stored in memory to let the light either pass or being deflect. Assuming the same pixel resolution (2 megapixel or 4K), readily available DMDs are characterized by at least 2 orders of magnitude faster (tens of kilohertz) settling speed compared to

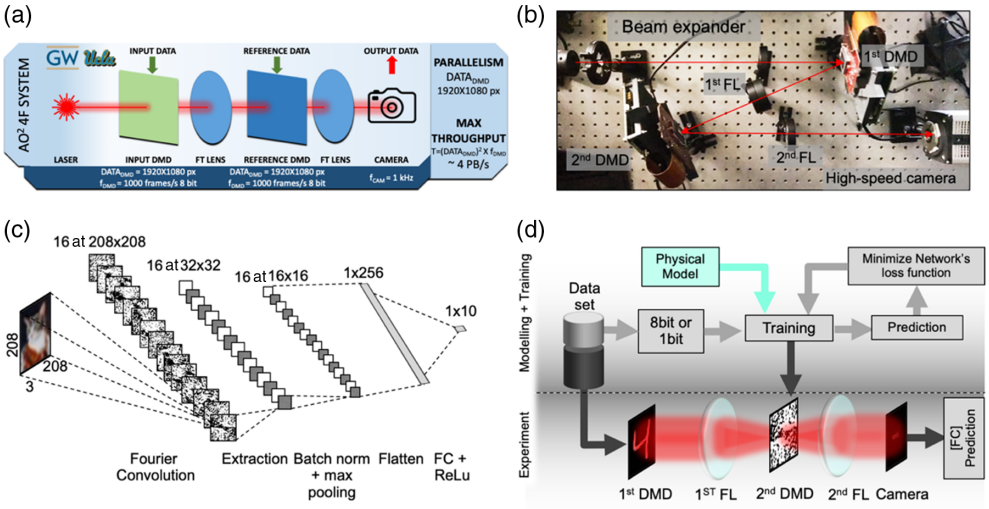

Fig. 1. Amplitude-only Fourier neural network. (a) Schematic representation of a 4 f system based on DMDs. The amplitude of a low-power light source is modulated according to a pattern (input data). The image so generated is Fourier transformed and multiplied with a reference data in the Fourier plane of a 4 f system, affecting only its amplitude. The result of the product is inverse transformed, and the square of its intensity is imaged by the camera showcasing the same spatial resolution (pixel size and pitch) of the DMDs. (b) Experimental implementation of the amplitude-only Fourier filter based on a DMD 4 f system. (c) CNN structure for CIFAR 10 dataset. The optical amplitude-only Fourier filter is used as a convolution layer, with the subsequent layers realized electronically. The kernels obtained during physically meaningful training are loaded in the second DMD. After a convolution layer a nonlinear thresholding is applied to the output (rectified linear unit function) and are pooled together. A flatten layer collapses the spatial dimensions of the output into the channel dimension to which follows a fully connected layer and a nonlinear activation function. (d) Flow chart of the training process. Physical model of the amplitude-only Fourier filter layer is used for training the entire CNN. (c) Obtaining the weights for the kernel to be loaded in the second DMD of the convolution layer. Experimentally obtained results of the amplitude-only Fourier filtering are fed to the FC layer for performing the final prediction on unseen data.

SLMs (tens of hertz), making them a promising platform for optical computing, thus the subject of this study.

In our optical engine [Fig. 1(b)], collimated low-power laser light (633 nm, He–Ne Laser) is expanded to uniformly interest the entire active area of the first DMD in the object plane, which, by independently tilting each micromirror of its array according to a pre-loaded image, defines the input image (feature map). The DMD in the object plane is oriented with a 12° tilting angle with respect to the normal incidence and rotated in-plane by 45°. Light reflected from the DMD is Fourier transformed passing through the first Fourier lens at one focal length, f, apart from the first DMD in the object plane. The pattern in the second DMD, with specular orientation with respect to the first one, acts as a spatial mask in the Fourier plane, opportunely selecting the spatial frequency components of the input image. The frequency filtered image is inverse Fourier transformed into the real space by the second Fourier lens and imaged by a high-speed camera [Fig. 1(b)]. Both FT transformation steps are performed entirely passively, i.e., zero-static power consumption, which is in stark contrast to performing convolutions as dot product multiplications in electronics [5].

On the system level, a computer loads the input image as well as the kernel (1920 × 1080, 8 bit, 1000 Hz) stored in its memory to the DMDs by means of a HDMI cable or directly generated through a field programmable gate array (FPGA) (Virtex 7), which connects to the digital light processing (DLP) boards (Texas Instrument) of the two DMDs through a serial connection, aiming to reduce the latency in providing the signals and allowing for processing while streaming data. Consequently, the AO Fourier filtered images are detected with a charge-coupled device (CCD) camera (1000 frames/s with 8 bit resolution) connected through PCI-express to a unified system interface which can store the data or process it implementing other NN tasks, such as max pooling, activation function, and fully connected (FC) layer. For emulating deeper neural networks which comprise multiple layers, the resulting image could be potentially loaded in the first DMD (see more details in Section 1 of Supplement 1.

Considering the abovementioned specifications, the system leverages (1) the vast parallelism given by the high pixel resolution of the camera and DMDs (2 megapixels); (2) inherent and completely passive operations due to the wave nature of the optical radiation, which allows for passive Fourier transforming exploiting lenses (Fresnel's integral) and pixelwise multiplication in the Fourier plane (Huygens' principle); (3) order of magnitude faster update rates compared to SLMs based on liquid crystals; thus (4) enabling a nominal throughput equivalent to 4 peta operations per second performed by space domain convolution operations (sliding window), with a resolution given by the DMDs (1920 × 1080 at 8 bit), updating at a frequency of ∼1 kHz and with a CCD camera acquisition frame rate of 1 kHz. It is worth stressing that, unlike other implementations [13,16] in which the kernels are fixed phase masks (diffractive elements or optical transparencies) and cannot be adjusted after training without physically replacing it, in our convolutional layer both feature maps and kernels can be updated at the same high rate (10 kHz). This can be particularly advantageous for emulating on the same hardware, deeper CNN architecture, which comprises multiple convolutional layers, in which batch normalization and max pooling are performed in the electrical domain. Notice that our convolutional layer already provides a straightforward nonlinearity (threshold) without the need of all optical nonlinearities as proposed by other schemes [23], which provides similar effects of a rectified linear unit (ReLU)

[24]. In detail, after the linear operation computed in the spatial frequency filtering (convolution) executed by the 4 f system, at the image plane, the electric field intensity associated to light is squared (x^2 function) when detected by the camera. Moreover, we demonstrate that, for our network architecture and dataset, additional nonlinearities do not provide any particular benefit (Supplement 1, Section 10). An all-optical nonlinearity in combination with this Fourier-optical CNN approach will be reported elsewhere.

The proposed AO-FF could be particularly useful in systems in which the input images are already encoded in a coherent radiation (first · DMD is absent). More in detail, if the inputs are already in the optical domain, the system which is opportunely trained using the proposed algorithm, can behave as a passive filter and therefore operate in real time, with execution time limited only by the integration time of the camera. The AO-FF can detect images within images (such as in steganography and optical illusions as shown in Section 2 of Supplement 1), demonstrating an immediate use in augmented visual perception or in classification of complex pattern, such as in iris recognition 8 bit scans or pattern recognition in LIDAR application.

Interestingly, spatial frequency filtering performed by a DMD is insensitive to the phase information. It is well established that full-field control could be achieved but here it is not desired. In 1963, in fact, Van der Lugt proposed a way to achieve plane frequency mask which retains effective phase and amplitude control in spite of using just absorption patterns [7], by exploiting Fourier holograms of the input image. Other full-field spatial control can be achieved through several interferometric schemes [25], such as Rayleigh or Mach–Zehnder interferometer, Lee holograms [26], superpixel [27], and more recent high-precision methods [28] and NN-based holographic reconstruction [29]. The full control over the optical field, while being advantageous in terms of image processing, comes at a cost of (1) increased complexity of the system, requiring additional optics and cumbersome alignments; and (2) reduction of the total dimension of the phase mask or need for corrective measurements and consequent drop of the overall parallelism. For these reasons, unlike other demonstrations [30], we deliberately decide to train the CNN to account for the information loss related to phase, and for the imprecise reconstruction of the images, while performing convolutions.

The designed CNN architecture consists of a single convolution layer in which sets of kernels are convolved with the input images. The convolutional layers are usually intercalated by pooling layer, which reduces the matrix dimensionality followed by nonlinear thresholding. Typical multilayer CNNs comprise layers of convolutional nodes followed by layers of fully connected nodes. Here, we use our experimental optical AO Fourier convolutional layer, whose output is pooled together, followed by a fully connected layer and nonlinear thresholding, both performed electronically. The convolutional layer has 16 nodes and each convolutional node uses a 208 × 208 kernel. The kernel parameters comprise the weights that are learned during the training procedure [Fig. 1(c)]. The CNN is trained using PyTorch, which is agnostic to the optics hardware. Therefore, it uses a set of functions, which exhaustively describe the Fourier convolution layer in order to accurately simulate the physical system. We adopt the concept of fast Fourier transform (FFT)-based Fourier domain training [31], together with the refined hardware model to accurately simulate the complete process and learn the kernel weights during training. The kernel values, which are the learnable parameters of the convolutional layer, are initialized directly in the Fourier domain. By doing so, the kernels do not need to be transformed into the Fourier domain such as required in [32,33], which matches our physical model well. For fully utilizing the maximum update speed of the DMD we restrict the kernel values to be real and binary; therefore, in the training a custom binarization step is needed. The CNN is trained using two classic datasets for image recognition to demonstrate the learning capability of this system as well as benchmarking it, namely the MNIST dataset of handwritten digits and CIFAR-10, a more challenging image classification problem. The trained kernel is used as an input pattern in the free-space 4 f system and the results of the convolutions are used for validating the physical model and for eventual successive training of the FC NN [Fig. 1(d)].

For obtaining a correct training and consequently highly accurate inference when performing convolution using the optical hardware, the physical model embedded into the training phase needs to accurately describe the coherent optical engine including its analog computing approximations and inaccuracies (for more details, see Section 3 of Supplement 1).

In order to validate the model and compare the results with the experimental realization of the optical engine, at first, we filter, by way of example, the 8 bit image of the GWU mascot (the Colonial), using different spatial frequency filters (Supplement 1). The results of the convolution obtained through the model and the experimental realization highlight a qualitative and quantitative agreement obtaining high (>0.7 for all the kernel except low-pass filter) structural similarity (SSIM), which is related to the image degradation as perceived change in structural information, and extremely low absolute errors, showcased by <0.1 root mean square error (more details in Section 4 of Supplement 1).

Leveraging on the massive amount of parallelism available in optical hardware (2 megapixels), the AO Fourier-based convolutional layer can be further parallelized if the input images (208 × 208 pixel) are smaller compared to the resolution offered by the DMD and the camera. In our experiment, we tiled in the input plane and batch process up to 46 images using the same kernel in the Fourier plane. Alternatively, the same input can be simultaneously filtered by multiple kernels; in this case, the Fourier transformed image is directed to different (nonoverlapping) portions of the DMD (or different DMDs) in the Fourier plane using opportune beam splitters, array of mirrors, and well-dimensioned microlens arrays. Ultimately each product is inverse Fourier transformed (using a second lenslet array) and imaged by different sensors. The filtered images can be integrated by the same sensor, performing dimensionality reduction. For additional information regarding the experimental implementation of the parallelization schemes, see Section 5 of Supplement 1.

After the model validation and establishment of parallelization schemes, to demonstrate the performance of the all-optical Fourier neural network (AO-FNN), we first trained the processor as an image classifier, performing automated classification of handwritten digits (MNIST). For this task, we train a one-layer convolutional layer, followed by a FC layer, with 55,000 images (5000 validation images) from the MNIST handwritten digit database. The input digits are encoded as amplitude and the network is trained to obtain the kernels (16, 208 × 208 binary images) to multiply in the Fourier plane, to be fed to the second DMD [Fig. 2(a)]. More details on the training are provided in Section 6 of Supplement 1.

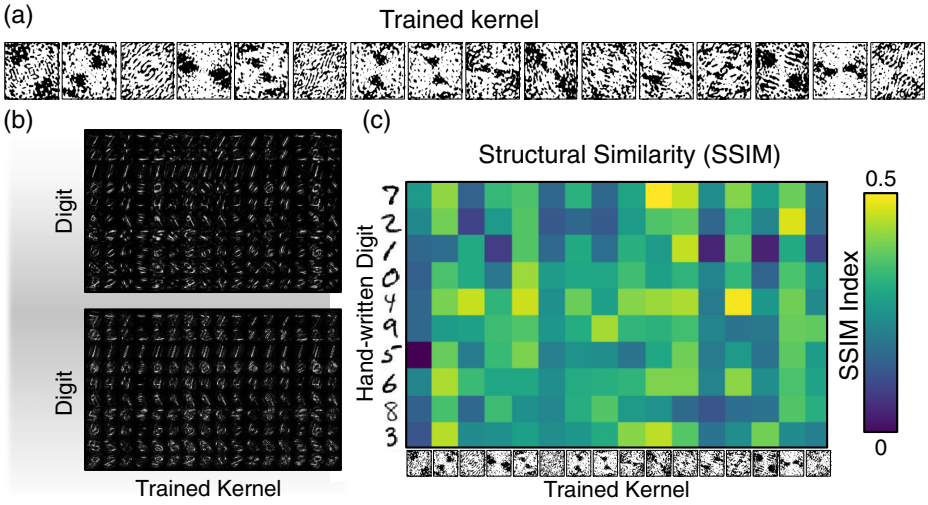

Fig. 2. Experimental testing of MNIST classifier. (a) Kernel obtained during training of the Fourier neural network for the classification of handwritten digits (MNIST dataset). (b) Output result of the emulated and experimental implementation of the first layer for different kernels (x axis) and input images (y axis). (c) Structural similarity map which compares the output obtained experimentally and those obtained during emulation for different digits (y axis) and kernels (x axis). We used the experimental output for training only the fully connected layer in order to compensate the discrepancies and improve the accuracy of the inference (see Visualization 1).

After the training, the network was blind tested, adopting the obtained kernel, using unseen images from the MNIST test dataset (not used as part of the training/validation), achieving 98% classification accuracy (Table 1). At this stage, for validating the hardware implementation, we perform convolutions between the kernels and unseen feature maps using the optical engine. The results of the emulated and experimental convolution layers are compared in terms of transformed feature maps and classification accuracy. Since our simulation model already considers some nonidealities of the optical hardware, the convolution results of the hardware implementation match the simulation results quite well qualitatively; their shapes are almost identical [Fig. 2(b)]. Although the match is not perfect quantitatively, highlighted by a lower SSIM [Fig. 2(c)]. This is due to several concurring factors including (a) small misalignment in the optical setup, (b) model which takes into account unphysical reflection of grid boundaries, and (c) nonideal camera dynamic range. The exact pixel values of hardware results differ from the simulation results; thus, if the convolution results obtained using the optical hardware are fed into a fully connected layer, whose weights are trained using simulation results, the actual classification accuracy will be significantly affected (92%). However, the Fourier kernel weights still bear the same representative information as the simulation model, and that the fully connected layer weights need to be updated to fit the hardware convolution results, thus compensating for the quantitative discrepancies between the model used for training and hardware implementation. Therefore, we implemented an ulterior fine-tuning process, which uses the hardware convolution results to retrain the fully connected weights of the layer with a reduced number of training samples. In detail, we perform the fine-tuning which utilizes the existing knowledge learned by the simulation model from the full training set and learns a mapping from the experimental results toward simulation results and compensates for it (Section 8). This approach proves to be particularly useful and the tuned hardware results accuracy shows a significant improvement (98%) compared with the one without fine-tuning (92%). Moreover, this fine-tuning approach which compensates from hardware-to-model discrepancies can be used if the optical engine is processing data in harsh environment conditions, for applications such as superresolution on object detection performance in satellite imagery, which can cause random misalignments.

For a more complex dataset, such as CIFAR-10, which comprises color images of 10 classes, with 6000 images per class, the inference accuracy for the simulated model is 62%, which is also close to the regularly used space-domain convolution model with full bit-precision, for similar neural network architecture (one conv. layer) implemented in different technology, such as one-layer electronic CNN or phase-only $4f$ schemes (accuracy of 51%). This is a promising result, since we show that in simulation our network with binarized kernel weights is able to obtain a (near) similar level of accuracy as normal space domain convolution using full precision features (32 bit). This can be explained by the effectiveness of the training of the $4f$ system, as well as the fact that there are more learnable parameters in the Fourier convolution, due to the larger kernel size compared with the space convolution version (more details in Section 11 of Supplement 1).

The ulterior degrees of freedom provided by the optical engine are considered to be "free" since the convolution time in the optical system does not depend on the size of the kernel as long as the size is within the DMD's resolution. After fine-tuning using a contained number (5000) of hardware results, the classification accuracy is a 54%, which is respectable given that it represents close to 90% from the nominal achievable results (Table 1).

To provide some details regarding the efficiency and performance of this novel computing scheme based on 2 megapixel DMDs, the AO-FF can perform convolutions between large matrices, in transform calculations, 10 times faster than a Nvidia P100 graphics card, commonly employed for high-performance computation, and more than 2 orders of magnitude faster than architectures which exploit SLMs, while consuming similar power. In terms of efficiency [Fig. 3(a)], the largest portion of the energy consumption and processing time of our optical engine comes from the signal transduction step, from digital electronics to optical domain and vice versa. In our optical system, the processing time for performing an 8 bit convolution is given by the sum of all delays, including the generation of the patterns (DMDs), time-of-flight of the

Table 1. Result of Normal Space Domain Convolution, Our Fourier Convolution Simulation Model, Hardware Model with and without Fine-Tuning[a]

Model	MNIST (%)	CIFAR (%)
Space-domain convolution (full precision)	98	63
Simulation model (Fourier convolution)	98	62
Hardware model (without fine-tuning)	92	25
Hardware model (with fine-tuning)	98	54

[a]For more details on simulation results, see Section 9 of Supplement 1.

photons through the optical setup, detection by the CCD camera (camera), and ultimately being transmitted for subsequent software processing. For a 2 megapixel 8 bit input and kernel images, the largest contribution latency is given by the camera acquisition time, followed by the DMD switching speed. The propagation time is negligible since, considering the $4f$ distances into play and the optical tools, it amounts for few nanoseconds. Whereas, the acquisition time of the high-speed camera is a function of the resolution of the image to be detected and represents the bottleneck of this current implementation. A higher-speed camera can ameliorate the processing time by a factor of 2, keeping the same DMD speed and resolution.

Looking at the future potential of this $4f$-based hybrid accelerator paradigm, developments of faster and higher-resolution spatial modulators and high-speed detection mechanisms are crucial to the advancement toward the implementation of intelligent functionalities [Fig. 3(b)]. For instance, higher-resolution DMDs (4 K resolution) and cameras would lead to an even increased parallelism (16 times the current throughput) compared to our prototype. Interestingly, at the research level, the analog version of MOEMS can reach high modulation speed (~ 10 MHz) without trading off pixel resolution (~ 10 megapixels) [38]. Using the analog MOEMS, for spatially modulating the optical signal, in combination with the ultra-high-speed camera (MHz, > 4K resolution), for converting the filtered signal in the electric domain, could improve the throughput of the system by about 4 orders of magnitude. However, for this configuration, the electronic interface will be the bottleneck of the system, which has to be able to deliver the patterns and acquire data with an overall bandwidth of tens of ~ 100 tera operations per second. Nonetheless, our AO $4f$ optical processor demonstration paves the way to future realizations; for instance, exploiting emerging technology components such as micrometer-thin metalenses, gigahertz fast reprogrammable metasurfaces, and high-speed photodiode arrays would yield to highly competitive footprint, while augmenting the computation throughput up to exa-operations per seconds, without trading off in terms of power consumption. However, at the current stage, these components are still challenged in terms of matrix resolution and achievable phase contrast [17,18]. These devices necessitate materials and device configurations which can provide efficient light–matter interactions, CMOS compatibility, straightforward and stark tunability, and sufficient maturity to be scaled up.

3. CONCLUSIONS

In summary, we have demonstrated an amplitude-only electro-optic Fourier filter engine with high-speed programmability and throughput. The dynamic Fourier filtering is realized using digital micromirror devices, both in the object and Fourier plane of an optical $4f$ system. As a proof-of-principle demonstration, we constructed a neural network which uses, as convolutional layer, the electro-optical convolutional engine for classifying handwritten digits (MNIST) and color images (CIFAR-10). We trained the network off-chip, using a detailed physical model which describes

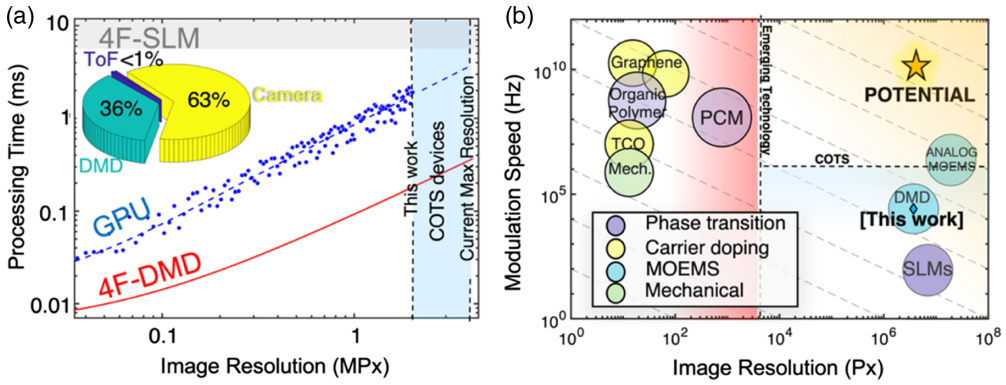

Fig. 3. Performance of the amplitude-only optical Fourier engine and its performance potential. (a) Comparison of total processing time for performing a convolution as a function of the image (matrix) resolution (expressed in megapixels) comparing the amplitude-only Fourier filter (red solid line) to the P100 Nvidia GPU (blue-dashed line fitting, experimental data dots) and a $4f$ system based on spatial light modulators (gray line). Here, we consider the convolution between two images (input and kernel) sharing the same pixel resolution expressed in megapixels. The 2 megapixel mark set the current maximum resolution of the DMD of this experimental realization but does not represent a technological limit. Pie chart illustrates the breakdown of the latency for the DMD-based $4f$ system when performing convolution. The overall latency consists of the DMD operation time (switching speed of the mirrors—green slice), camera integration time (yellow slice), and time of flight of the photon in the optical setup (violet slice). (b) Programmable electro-optic spatial light modulator grouped according to the functioning principle define processor performance defined by matrix size-speed-product (gray iso-performance lines). Exemplary, the 100× improvement over an SLM-based system (e.g., Optalysys) is a direct function of matrix size and update rate: carrier doping (Graphene [34,35], TCO [36]), phase change (PCM [37], Organic Polymer [18], LCOS-SLM Gaea-2), MOEMS (Texas Instruments: 2MPx-DLP9000 and 4 K-DLP660TE, Analog MOEMS [38]), and electro-mechanical [39], which can contemporary increase the throughput and lower latency of the proposed $4f$ system. The plot is tripartite into emerging technologies, COTS devices, and potential hardware with GHz-fast, million-pixel electro-optic devices which can spatially modulate light for the next-generation information science and sensing.

the electro-optical system and its nonidealities, such as optical aberrations and misalignments. After experimentally validating the model and retraining the following fully connected layer to compensate for value discrepancies, we obtained a classification accuracy of 98% and 54% for MNIST and CIFAR-10, respectively, with a throughput up to 1000 convolutions per seconds between two 2 megapixel images, which is 1 order of magnitude faster than the state-of-the-art GPU. Additionally, our scientific contribution emphasizes that the information loss and inaccuracies deriving from neglecting the phase of the optical wavefront can be compensated for by the degree of robustness provided by neural network training, which yields intelligent classification, at as high accuracy as the one obtained by phase-only optical engine, while featuring 2 orders of magnitude faster programmability. The system can also be used to filter images of smaller resolution in parallel, and by exploiting *ad hoc* electronic I/O interface, emulate deeper neural networks reaching high number of connections and millions of neurons. This paradigm and hardware implementation of the optical engines for artificial neural networks is a promising alternative to other machine learning architecture since they can avail parallel computing capability and power efficiency inherent to optical systems. Our results, reported for different inference tasks, indicate the potential that our intelligent information processing scheme could open new perspectives of flexible and compact platforms which could be transformative for diverse applications, ranging from image analysis to image classification and superresolution imaging on unmanned aerial vehicles, and may also enable high-bandwidth free-space communication in data centers, intelligently pre-processing data locally at the edge of the network.

Funding. Army Research Office (W911NF1910468); Office of Naval Research (N00014-19-1-2595).

Acknowledgment. We thank Prof. Aydin Babakhani, Prof. Seth Bank, Prof. Tarek El-Ghazawi, Prof. David Pan, and Prof. Chee Wei Wong of the "Photonic Convolutional Processor for Network Edge Computing" project for the insightful discussions. V. J. S. is supported by the Advanced Computing Program (ACI) under the Army Research Office.

V. J. S. and M. M. envisioned the idea of an amplitude-only Fourier convolutional engine for deep learning, V. J. S. and P. G. acquired the funds and supervised the project. M. M. developed the relevant theories and analyses for the project. M. M. and Z. H. designed the experimental setup and conducted the free-space experiments. R. C. and H. D. supported the system's high-speed data input-output. S. L. designed and trained the amplitude-only Fourier neural network and performed the relevant tests and benchmarking. M. M., P. G., S. L. P. H., J. G., and V. J. S. discussed the results and contributed to the understanding, analysis, and interpretation of the results. All authors contributed to writing the manuscript.

Disclosures. M. M. and V. J. S. (P); H. D. Optelligence LLC (E, I); V. J. S. Optelligence LLC (I).

See Supplement 1 for supporting content.

REFERENCES

1. Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "LipNet: end-to-end sentence-level lipreading," arXiv:1611.01599v2 (2017), p. 13.
2. T. Simonite, "Google's AI wizard unveils a new twist on neuralnetworks," in WIRED, 1 November 2017.
3. V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," Nature **518**, 529–533 (2015).
4. X. Li, G. Zhang, H. H. Huang, Z. Wang, and W. Zheng, "Performance analysis of GPU-based convolutional neural networks," in *45th International Conference on Parallel Processing (ICPP)* (IEEE, 2016), pp. 67–76.
5. D. Li, X. Chen, M. Becchi, and Z. Zong, "Evaluating the energy efficiency of deep convolutional neural networks on CPUs and GPUs," in *IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom) (BDCloud-SocialCom-SustainCom)* (IEEE, 2016), pp. 477–484.
6. H. M. Ozaktas and H. Urey, "Space-bandwidth product of conventional Fourier transforming systems," Opt. Commun. **104**, 29–31 (1993).
7. A. V. Lugt, "Signal detection by complex spatial filtering," IEEE Trans. Inf. Theory **10**, 139–145 (1964).
8. "Optical neural computers," https://www.scientificamerican.com/article/optical-neural-computers/.
9. A. V. Lugt, "Coherent optical processing," Proc. IEEE **62**, 1300–1319 (1974).
10. C. S. Weaver and J. W. Goodman, "A technique for optically convolving two functions," Appl. Opt. **5**, 1248–1249 (1966).
11. Y. Chen, "4f-type optical system for matrix multiplication," Opt. Eng. **32**, 77–79 (1993).
12. D. Psaltis, D. Brady, X.-G. Gu, and S. Lin, "Holography in artificial neural networks," Nature **343**, 325–330 (1990).
13. X. Lin, Y. Rivenson, N. T. Yardimci, M. Veli, Y. Luo, M. Jarrahi, and A. Ozcan, "All-optical machine learning using diffractive deep neural networks," Science **361**, 1004–1008 (2018).
14. "Optalysys completes 320 gigaFLOP optical computer prototype, targets 9 petaFLOP product in 2017 and 17 exaFLOPS machine by 2020—NextBigFuture.com," https://www.nextbigfuture.com/2015/05/optalysys-completes-320-gigaflop.html.
15. T. Lu, S. Wu, X. Xu, and F. T. S. Yu, "Two-dimensional programmable optical neural network," Appl. Opt. **28**, 4908–4913 (1989).
16. J. Chang, V. Sitzmann, X. Dun, W. Heidrich, and G. Wetzstein, "Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification," Sci. Rep. **8**, 1–10 (2018).
17. J. Sun, E. Timurdogan, A. Yaacobi, E. S. Hosseini, and M. R. Watts, "Large-scale nanophotonic phased array," Nature **493**, 195–199 (2013).
18. A. Smolyaninov, A. El Amili, F. Vallini, S. Pappert, and Y. Fainman, "Programmable plasmonic phase modulation of free-space wavefronts at gigahertz rates," Nat. Photonics **13**, 431–435 (2019).
19. J. L. Horner and P. D. Gianino, "Phase-only matched filtering," Appl. Opt. **23**, 812–816 (1984).
20. J. M. Kahn and D. A. B. Miller, "Communications expands its space," Nat. Photonics **11**, 5–8 (2017).
21. J. Goodman, *Introduction to Fourier Optics*, 3rd ed. (Roberts and Company, 2005).
22. A. V. Oppenheim and J. S. Lim, "The importance of phase in signals," Proc. IEEE **69**, 529–541 (1981).
23. Y. Zuo, B. Li, Y. Zhao, Y. Jiang, Y.-C. Chen, P. Chen, G.-B. Jo, J. Liu, and S. Du, "All-optical neural network with nonlinear activation functions," Optica **6**, 1132–1137 (2019).
24. S. Chen, X. Wang, C. Chen, Y. Lu, X. Zhang, and L. Wen, "DeepSquare: boosting the learning power of deep convolutional neural networks with elementwise square operators," arXiv:1906.04979 [cs] (2019).
25. M. Mirhosseini, O. S. Magaña-Loaiza, C. Chen, B. Rodenburg, M. Malik, and R. W. Boyd, "Rapid generation of light beams carrying orbital angular momentum," Opt. Express **21**, 30196–30203 (2013).
26. W.-H. Lee, "Binary synthetic holograms," Appl. Opt. **13**, 1677–1682 (1974).

27. S. A. Goorden, J. Bertolotti, and A. P. Mosk, "Superpixel-based spatial amplitude and phase modulation using a digital micromirror device," Opt. Express **22**, 17999–18009 (2014).
28. L. Liu, Y. Gao, and X. Liu, "High-precision joint amplitude and phase control of spatial light using a digital micromirror device," Opt. Commun. **424**, 70–79 (2018).
29. Y. Rivenson, Y. Zhang, H. Günaydın, D. Teng, and A. Ozcan, "Phase recovery and holographic image reconstruction using deep learning in neural networks," Light Sci. Appl. **7**, 17141 (2018).
30. E. G. Paek, J. R. Wullert, A. Von Lehmen, J. S. Patel, A. Scherer, J. Harbison, H. J. Yu, and R. Martin, "VanderLugt correlator and neural networks," in *Conference Proceedings., IEEE International Conference on Systems, Man and Cybernetics* (1989), Vol. **2**, pp. 408–414.
31. H. Pratt, B. Williams, F. Coenen, and Y. Zheng, "FCNN: Fourier convolutional neural networks," in *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, M. Ceci, J. Hollmén, L. Todorovski, C. Vens, and S. Džeroski, eds. (Springer International Publishing, 2017), pp. 786–798.
32. T. Abtahi, A. Kulkarni, and T. Mohsenin, "Accelerating convolutional neural network with FFT on tiny cores," in *IEEE International Symposium on Circuits and Systems (ISCAS)* (2017), pp. 1–4.
33. M. Mathieu, M. Henaff, and Y. LeCun, "Fast training of convolutional networks through FFTs," arXiv:1312.5851 [cs] (2014).
34. Y. Yao, R. Shankar, M. A. Kats, Y. Song, J. Kong, M. Loncar, and F. Capasso, "Electrically tunable metasurface perfect absorbers for ultrathin mid-infrared optical modulators," Nano Lett. **14**, 6526–6532 (2014).
35. B. Zeng, Z. Huang, A. Singh, Y. Yao, A. K. Azad, A. D. Mohite, A. J. Taylor, D. R. Smith, and H.-T. Chen, "Hybrid graphene metasurfaces for high-speed mid-infrared light modulation and single-pixel imaging," Light Sci. Appl. **7**, 51 (2018).
36. R. Amin, R. Maiti, Y. Gui, C. Suer, M. Miscuglio, E. Heidari, R. T. Chen, H. Dalir, and V. J. Sorger, "Sub-wavelength GHz-fast broadband ITO Mach–Zehnder modulator on silicon photonics," Optica **7**, 333–335 (2020).
37. H.-S. Ee and R. Agarwal, "Electrically programmable multi-purpose non-volatile metasurface based on phase change materials," Phys. Scr. **94**, 025803 (2019).
38. J.-U. Schmidt, U. A. Dauderstaedt, P. Duerr, M. Friedrichs, T. Hughes, T. Ludewig, D. Rudloff, T. Schwaten, D. Trenkler, M. Wagner, I. Wullinger, A. Bergstrom, P. Bjoernangen, F. Jonsson, T. Karlin, P. Ronnholm, and T. Sandstrom, "High-speed one-dimensional spatial light modulator for laser direct imaging and other patterning applications," Proc. SPIE **8977**, 89770O (2014).
39. N. I. Zheludev and E. Plum, "Reconfigurable nanomechanical photonic metamaterials," Nat. Nanotechnol. **11**, 16–22 (2016).