Advanced Photonics Congress (IPR, Networks, NOMA, PVLED, SPPCom) © OSA 2021

Fourier Optical Convolutional Neural Network Accelerator

Mario Miscuglio[1], Zibo Hu[1], Shurui Li[2], Puneet Gupta[2], Hamed Dalir[1,3], Volker J. Sorger [1,3]*

[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, Los Angeles, CA, 90095, USA
[3] OPTELLIGENCE Company, 300 Delaware Ave, Wilmington, 19801, DE, USA
*Correspondence to E-mail address: sorger@gwu.edu and sorger@optelligence.co

Abstract: Here we report a massively-parallel Fourier-optics convolutional processor accelerated 160x over spatial-light-modulators using digital-mirror-display technology as input and kernel showing an MNIST and CIFAR-10 accuracy of 96% and 54%, respectively.

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 photonic integrated circuits [1-5] and free-space optics [6] 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 neural network 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).

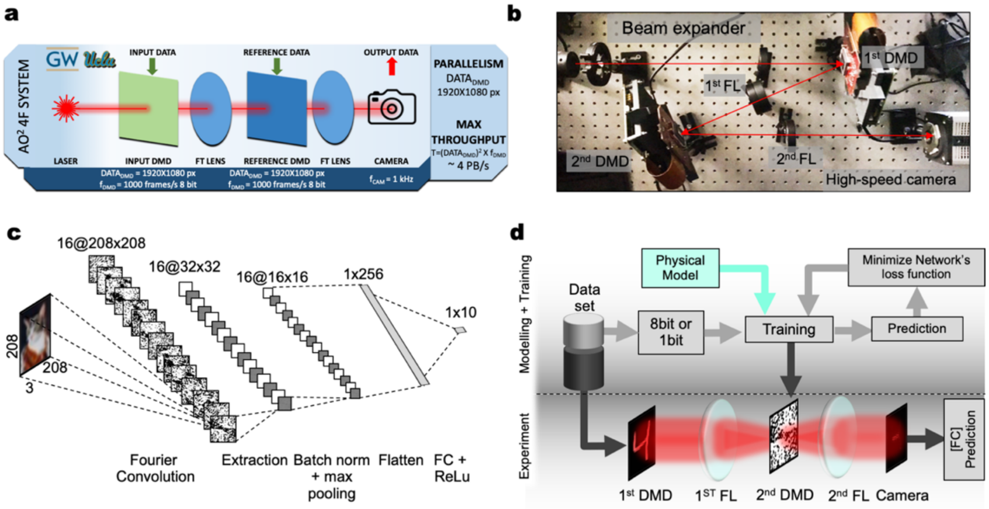

Figure 1. Amplitude only Optical Fourier Neural Network. a, Schematic representation of a 4f system based on a Digital Micromirror Devices (DMDs). The amplitude of a low power light source is modulated according to a pattern (input data). The image so generated is Fourier transformed and multiplied with a reference data in the Fourier plane of a 4f 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 4F system. **c**, Convolutional Neural Network (CNN) structure for CIFAR 10 dataset. The optical Amplitude only Fourier filter is used as convolution layer, with the subsequent layers realized electronically. The kernels obtained during physically meaningful training are loaded in the 2nd 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. Obtaining the weights for the kernel to be loaded in the 2nd 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. **d**, Training of the CNN involving a physical model.

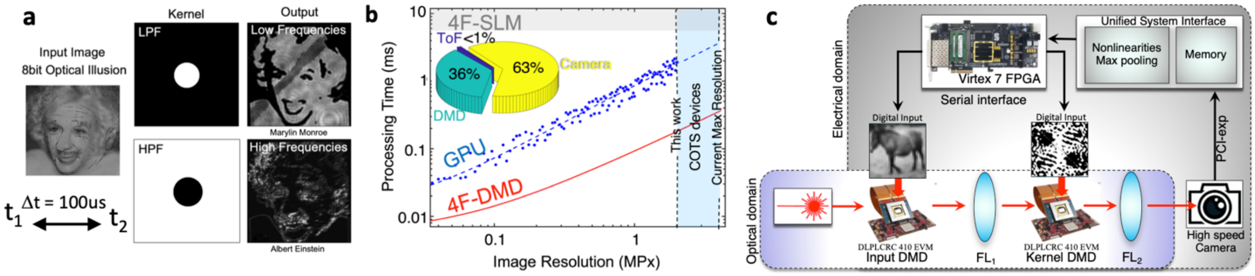

Figure 2. **a**, This fast switchable Kernel allows for rapid filtering; mock-demonstration showing optical illusion. **b**, Performance of the amplitude-only optical Fourier engine and its performance potential. a Comparison of total processing time for performing a convolution as function of the image (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 (grey line). Here, we consider the convolution between two images (input and kernel) sharing the same pixel resolution expressed in MPx. The 2MPx 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). **c**, System schematic of Fourier-optical system deploying DMDs, high-speed camera, and an FPGA for data I/O. Future system enhancements include photonic DACs [7], and/or GHz-fast switching PIC technology based on modulators [8], or data or kernel multiplexing techniques [9].

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.

The system can be used for rapid-kernel switching of filtering and inference applications (**Fig. 2a**). This work has two far-reaching contributions; (i) scientifically it challenges the assumption that phase-information is critically needed in optical image processors for machine-intelligence, and (ii) it experimentally demonstrates a prototype with 10x lower latency than the top-of-the-line GPU for the same matrix-size. Such performance is achieved by a the remarkable throughput of up to 1 P-OPS at a 8-bit resolution, or 10 P-OPS at 1-bit for complex data-sets of CIFAR-10, or MNIST, respectively (OPS = operations per second). As such, this work is heralding a new era of amplitude-only signal processing without scarifying accuracy (**Fig. 2b**). Such performance is enabled by our realization that neural networks are analog processors, which allow to map the system homomorphically onto photonic hardware – specifically utilizing the spatial parallelism of digital display technology enabling processing large (1,000x1,000) matrices in one single time step, unlike electronics where kernel striding slows the convolution-processing algorithm down. Exemplary, here we demonstrate this amplitude-only paradigm on performing a convolutional neural network; such networks are the 'bread-and-butter' of machine-learning systems since they enable the critical feature extraction from data and images, and constitute about 80% of a neural network. On a systems level (**Fig. 2c**), the I/O rate is becoming an question of concern, demanding electronic support. A digital-to-analog conversions may need to be considered, which is a power hungry step. Recent photonic DAC developments could streamline data conversions [7].

In conclusion, we introduced a massively parallel (2 million channel) optical convolutional neural network based on Fourier domain rapid frequency filtering. We demonstrate signal process capability with fast (~0.1ms) updating filters using a dual digital-mirror-display 4f-system approach. Offline trained kernels and performing inference classification tasks on CIFAR-10 dataset shows an 87% relative accuracy against full-precision.

References

[1] M. Miscuglio, et al. Roadmap on Material-Function Mapping for Photonic-Electronic Hybrid Neural Networks, *APL Mat.* 7, 100903 (2019).
[2] M. Miscuglio, V. J. Sorger "Photonic Tensor Cores for Machine Learning", Applied Physics Reviews 7, 031404 (2020)
[3] M. Miscuglio *et al.* All-optical nonlinear activation function for photonic neural networks, *Opt. Mat. Exp.* 8, 12, p. 3851, (2018).
[4] J. K. George, et al. Noise and Nonlinearity of Electro-optic Activation Functions in Neuromorphic Compute Systems, *Opt. Exp.* 27, 4 (2019).
[5] A. Mehrabian, et al. A Winograd-based Integrated Photonics Accelerator for Convolutional Neural Networks, IEEE J. Sel. Top. Qua. El. (2019).
[6] M. Miscuglio, Z. Hu, S. Li, J. K. George, et al. "Massively parallel amplitude-only Fourier neural network," Optica 7, 1812-1819 (2020)
[7] J. Meng, M. Miscuglio, J. K. George, A. Babakhani, V. J. Sorger "Electronic Bottleneck Suppression in Next - generation Networks with Integrated Photonic Digital-to-analog Converters" Adv. Phot. Research, 2, 2000033 (2021).
[8] R. Amin, R. Maiti, et al. "Broadband Sub-λ GHz ITO Plasmonic Mach-Zehnder Modulator on Silicon Photonics" OPTICA 7, 3, (2020).
[9] S. Liu, et al. "Channel Tiling for Improved Performance and Accuracy of Optical Neural Network Accelerators" arXiv preprint: 2011.07391.
[10] R. Amin et al. "ITO-based electro-absorption modulator for photonic neural activation function" APL Materials 7, 8, 081112 (2019).
[11] We acknowledge support from the Office of Navy Research under award number N00014-19-1-2595 of the Electronic Warfare program.