Massively-parallel Amplitude-Only Fourier Optical Convolutional 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], Volker J. Sorger [1,*]

[1]*Deptartment of Electrical and Computer Engineering, George Washington University, Washington DC, DC, USA*
[2]*Deptartment of Electrical and Computer Engineering, University of California, Los Angeles, CA, USA*
[3]*Department of Mechanical and Aerospace Engineering, George Washington University, Washington, DC, USA*
[4]*Optelligence LLC, Virginia, VA, USA*
**sorger@gwu.edu*

Abstract: Here we introduce a novel amplitude-only Fourier-optical processor paradigm and demonstrate a prototype system capable of processing large-scale ~(2,000x1,000) matrices in a single time-step and 100 microsecond-short latency, for accelerating machine-learning applications.

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 towards processor heterogeneity [1]. Task-specific accelerators [2] such as photonic tensor cores [3], and those based on free-space optics [4] bear fundamental homomorphism for massively parallel and real-time information processing given the wave-nature of light [5-7]. 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 ~(1,000 × 1,000) matrices in a single time-step and 100 microsecond-short latency (**Fig. 1**). 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 pixel-wise multiplications in the Fourier domain reaching peta operations per second throughputs. The required real-to-

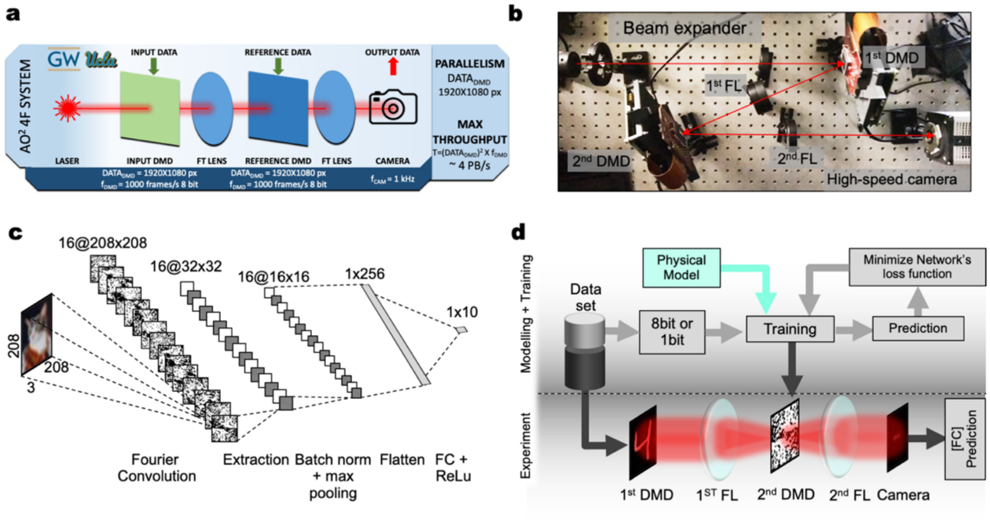

Figure 1. Amplitude only Fourier Neural Network. **a** Schematic representation of a 4f system based on a Digital Micromirror Devices (DMDs). The amplitude of a low power light source is modulated according to a pattern (input data). The image so generated is Fourier transformed and multiplied with a 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. (**c**), 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.

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 GPU and phase-based display technology by one and two orders of magnitude, respectively (**Fig. 2**). Training this optical convolutional layer on image classification tasks and utilizing it in a hybrid optical-electronic CNN, shows classification accuracy of 98% (MNIST) and 54% (CIFAR-10). Interestingly, the amplitude-only CNN is inherently robust against coherence noise in contrast to phase-based paradigms and features an over 2 orders of magnitude lower delay 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 trained, 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 towards near real-time decision making. This amplitude-only electro-optic Fourier filter engine prototype offers high-speed kernel programmability and data throughput (**Fig. 1&2**). 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 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 values discrepancies, we obtained a classification accuracy of 98% and 54% for MNIST and CIFAR-10, respectively, with a throughput up to 1,000 convolutions per seconds between two 2MP images, which is one 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 wave front can be compensated-for by the degree of robustness provided by neural network training, which yields intelligent classification, at high accuracy as the one obtained by phase-only optical engine, while featuring a 160x faster programmability. Future systems could be augmented with photonic DACs [8] for applications where data is present in the optical domain. We built, and tested a massively parallel (2 million channel) optical convolutional Fourier processor and 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 of this 1-layer convolutional processor shows an encouraging (-10%) accuracy against full-precision [9]. For funding support see Ref [10].

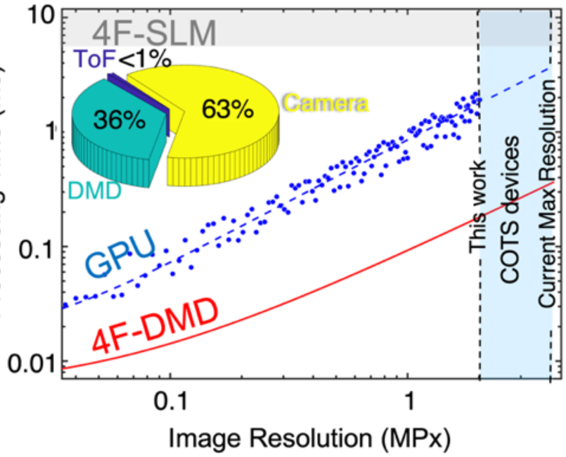

Figure 2. Performance of the amplitude-only optical Fourie engine and its performance potential. a Comparison of tota processing time for performing a convolution as function of the image (matrix) resolution (expressed in Megapixels) comparing the Amplitude-only Fourier Filter (red solid line) to the 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).

References

[1] S. Sun, et al. CLEAR: A Holistic Figure-of-Merit for Post- and Predicting Electronic and Photonic-based Compute-system Evolution, *Scientific. Reports*, 10:6482, 1-9 (2020).

[2] A. Mehrabian, et al. A Winograd-based Integrated Photonics Accelerator for Convolutional Neural Networks, *IEEE J. Sel. Top. Qua. El.* (2019).

[3] M. Miscuglio, V. J. Sorger, Photonic Tensor Cores for Machine Learning, *Applied Physics Reviews* 7, 031404 (2020).

[4] J. Chang, V. Sitzmann, X. Dun, W. Heidrich, and G. Wetzstein, *Scientific Reports* 8, 1–10 (2018).

[5] M. Miscuglio, et al. Roadmap on Material-Function Mapping for Photonic-Electronic Hybrid Neural Networks, *APL Mat.* 7, 100903 (2019).

[6] J. K. George, et al. Noise and Nonlinearity of Electro-optic Activation Functions in Neuromorphic Compute Systems, *Opt. Exp.* 27, 4 (2019).

[7] M. Miscuglio *et al.* All-optical nonlinear activation function for photonic neural networks, *Opt. Mat. Exp.* 8, 12, p. 3851, (2018).

[8] J. Meng, et al. Electronic Bottleneck Suppression in Next-generation Networks with Integrated Photonic Digital-to-analog Converters, *Advanced Photonics Research*, doi:10.1002/adpr.202000033 (2020).

[9] D. Li, et al. Evaluating the Energy Efficiency of Deep Convolutional Neural Networks on CPUs and GPUs, *IEEE BDCloud,* 477–484 (2016).

[10] We acknowledge support from the Office of Navy Research under award number N00014-19-1-2595 of the Electronic Warfare program.