Million-channel parallelism Fourier-optic convolutional filter and neural network processor

Mario Miscuglio[1], Zibo Hu[1], Shurui Li[2], Jiaqi Gu[3], Aydin Babakhani[2], Puneet Gupta[2], Chee-Wei Wong[2], David Pan[3], Seth Bank[3], Hamed Dalir[4], 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, Los Angeles, CA, 90095, USA*
[3]Department of Electrical and Computer Engineering, University of Texas, Austin, TX 78758, USA
[4]Omega Optics, Inc. 8500 Shoal Creek Blvd., Bldg. 4, Suite 200, Austin, Texas 78757, USA
**sorger@gwu.edu*

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

1. Introduction and Rationale

The rise of machine-learning has shown significant utility in i) deep-learning such as image and language classification, ii) nonlinear optimization such as in predictive control including target tracking, and iii) generative adversary networks. While electronic implementations provide utility, they often do not scale well in terms of inference delay or power consumption; for instance, a single game of AlphaGo Zero costs about $3,000 in electricity, and transistor scaling does not improve performance. Within the trend of hardware-specific accelerators driving a tend in computing heterogeneity, we envision future accelerators to include photonic process units (PPU), which utilize 1) one-shot (non-iterative) executions towards delivering $\boldsymbol{O}(1)$ process capability; 2) massive (10^6) parallelism such as form free-space digital light processing technology; 3) 'cheap' convolutions enabled by a 'free' Fourier transform performed by a lens; and 4) maturing foundry PDKs of high-performance photonic integrated circuits (PIC) [1-4]. Here we show a prototype of a Fourier-optics enabled 4f systems utilizing digital-mirror-display (DMD) technology for convolutional filtering and convolutional neural networks (CNN) (Fig. 1a).

2. Results and Discussion

The main idea is to perform massively parallelized vector matrix multiplications (VMM) in the Fourier domain as point-wise multiplications reducing the elsewise (e.g. GPU) $\boldsymbol{O}(n^2)$ multiplications to $\boldsymbol{O}(n)$; that is, any multiplication can be simplified by summing the digits of each n-digit factor and then multiplying. Furthermore, the in electronics costly Fourier transformation is here performed passively by a lens of this 4f Fourier processor (Fig. 1b), where in the Fourier-domain spatial frequency filtering is performed (~kernel). However, unlike spatial light modulators (SLM), which clock at 60 Hz rely on slow liquid crystal technology, here we deploy fast DMDs with the same spatial resolution but 10 kHz fast rates (~160x kernel speed-up, Fig. 1c). However, since DMDs are phase insensitive, phase information is not considered, but could be accounted for in a Vanderlugt lens system by sacrificing portion spatial resolution.

We built such a dual DMD system where the first DMD loads the image (or signal) and the second DMD performs the amplitude-based filter (Fig. 1c). As a test we loaded an image consisting of a differently rotated bar test-pattern into the processor. Then changing the kernel DMD, by rotating an exemplary (and ad-hoc selected) edge-detection kernel, we show that individual bar patterns can be selected. Such frequency processing can further be performed not only with images but also with bit-strings, RF-signals, or any other signal encoded in the optical domain.

In order to gain further insights into the accuracy resolution losses and performance of this 4f Fourier processor, we a) developed a physical accurate model of the 4f diffraction-based filtering of the 4f system, and b) perform offline kernel training and use the trained weights as Fourier kernels. Regarding the former, we develop a model that considers the phase-noise, limited numerical aperture, effects originating from aberration and diffraction of the lenses, and an accurate transfer function of the DMDs, which include dead-zones of the pixels of the DMD. Using this accurate description of our physical 4f system, we can now turn to the many offline training procedures in the field of machine-learning to train the kernels accurately (Fig. 1d). In brief a select an application-oriented open dataset (here MNIST, CIFAR-10) and perform a relatively standard gradient descent back-propagation-based training algorithm. For this, the network structure consists of 1x Fourier convolution layer, 1x fully connected (FC) layer (128-10) and use the ADAM optimizer. For the Fourier Conv. layer, the kernel is initialized with real numbers matching the

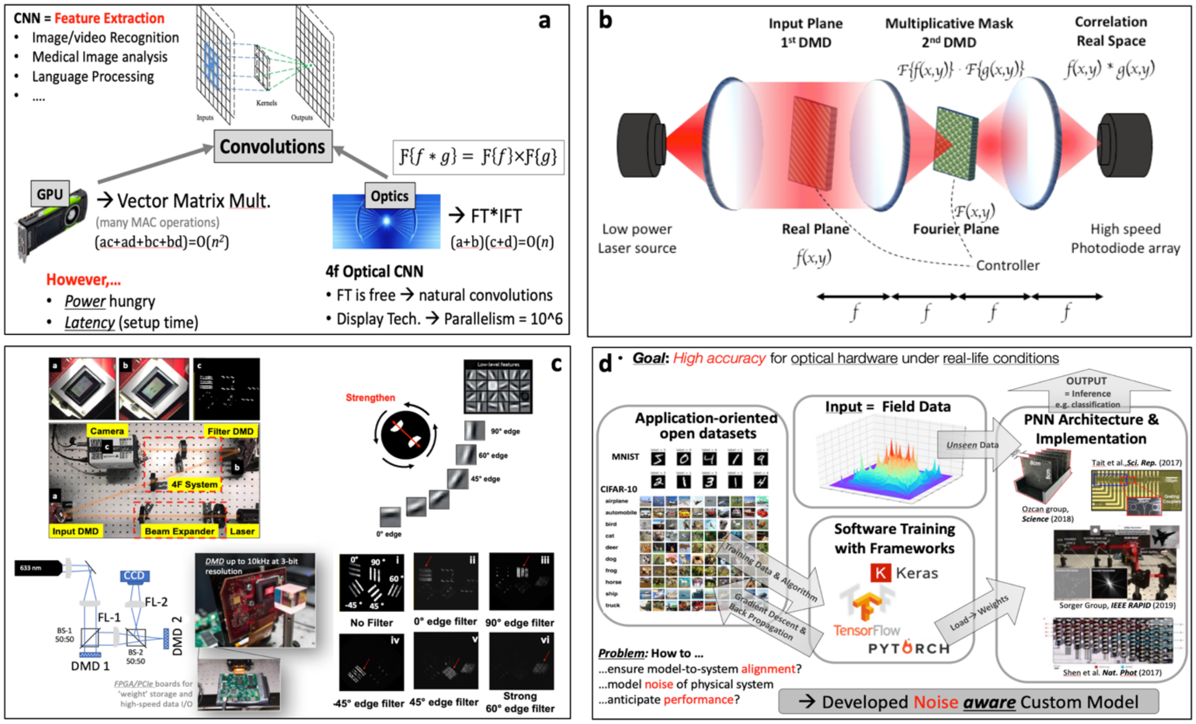
a
CNN = Feature Extraction
• Image/video Recognition
• Medical Image analysis
• Language Processing
• ....
Inputs
Kernels
Outputs
Convolutions
$\mathcal{F}\{f * g\} = \mathcal{F}\{f\}\times\mathcal{F}\{g\}$
GPU
→ Vector Matrix Mult.
(many MAC operations)
(ac+ad+bc+bd)=O(n²)
Optics
→ FT*IFT
(a+b)(c+d)=O(n)
However,...
• Power hungry
• Latency (setup time)
4f Optical CNN
• FT is free → natural convolutions
• Display Tech. → Parallelism = 10^6
b
Input Plane
1st DMD
Multiplicative Mask
2nd DMD
Correlation
Real Space
$\mathcal{F}\{f(x,y)\}\cdot\mathcal{F}\{g(x,y)\}$
$f(x,y) * g(x,y)$
Low power
Laser source
Real Plane
$f(x,y)$
$\mathcal{F}(x,y)$
Fourier Plane
Controller
High speed
Photodiode array
f f f f
c
Camera
Filter DMD
4F System
Input DMD
Beam Expander
Laser
Strengthen
Low-level features
90° edge
60° edge
45° edge
0° edge
633 nm
CCD
FL-2
FL-1
BS-1 50:50
BS-2 50:50
DMD 1
DMD 2
DMD up to 10kHz at 3-bit resolution
FPGA/PCIe boards for 'weight' storage and high-speed data I/O
No Filter
0° edge filter
90° edge filter
-45° edge filter
45° edge filter
Strong 60° edge filter
d • Goal: High accuracy for optical hardware under real-life conditions
Application-oriented open datasets
MNIST
CIFAR-10
airplane automobile bird cat deer dog frog horse ship truck
Input = Field Data
Unseen Data
OUTPUT = Inference e.g. classification
PNN Architecture & Implementation
Tait et al., Sci. Rep. (2017)
Ozcan group, Science (2018)
Sorger Group, IEEE RAPID (2019)
Shen et al. Nat. Phot (2017)
Training Data & Algorithm
Gradient Descent & Back Propagation
Software Training with Frameworks
Keras
TensorFlow
PYTORCH
Load → Weights
Problem: How to ...
...ensure model-to-system alignment?
...model noise of physical system
...anticipate performance?
→ Developed Noise aware Custom Model

Figure 1. a, The 'free' Fourier transform in optics provided by a lens enables natural; convolutions. Together with 10^6 parallel channels from display technology, high-end convolutional processors can be built in optics (100's TMAC/s). **b&c**, Schematic and prototype of 4f DMD -based processor enabling speed-up times over SLMs of 160x. **d**, Training concepts and rational for system accurate modeling.

DMD hardware and the size of the Kernel is an input image (32x32 pixels). The trained kernel weights then become the 2nd DMD cell values. The Fourier Conv. layer workflow is as follows: 1) Apply FFT on input image and transformation to Fourier domain. 2) Multiply the result with the Kernel in the Fourier plane. 3) Apply inverse Fourier transform to obtain result in space domain, then take the L2 norm (Matrix). 4) Feed the output to FC layer Kernel. Lastly, we test the convolution precision, namely, the inference accuracy, addressing the effects of a reduced bit-density (1-bit) of the DMD. We then test i) full and ii) 1-8 bits precision but restrict the kernel weights range to 0-1 (for the 1-8 bits) matching the DMD's bit-resolution. For the 1-bit case, negative = '0', positive = '1' (binarization). For other precisions, quantization is used (each forward pass, adjust weights). The result for this single Conv. layer network suggests that by initializing the kernel in Fourier domain directly using real number does not impact the learning result (at least for this single-layer version with MNIST or CIFAR). For the 1-bit mode, the result is not comparable to standard convolution, but still significantly better than only using 2 FC layers. However, for the 2-bit mode and above the result is close to standard convolution and above 4-bit precision the result is even better, probably due to the extra regularization effect of quantization as well as the 'simple' dataset. MNIST results (not-shown) give a 96.5% accuracy.

Structure	Standard conv (kernel size = 5)	Fourier conv (full precision)	Fourier conv (8-bit)	Fourier conv (6-bit)	Fourier conv (4-bit)	Fourier conv (2-bit)	Fourier conv (1-bit)	Full-connected layer (2 layers)
Result	63%	64%	65%	65%	63%	61%	54%	47%

Table 1. Inference results of a trained convolutional layer for the CIFAR-10 dataset showing comparable results to the full-precision kernel. The 160x time-accelerated optical 4f convolutional processor based on DMDs shows a 10% reduced accuracy.

In conclusion, here we discussed, built, and tested a massively parallel (2 million channel) optical convolutional Fourier processor. 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 of this 1-layer convolutional processor shows an encouraging (-10%) accuracy against full-precision [5].

3. 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 *et al.* All-optical nonlinear activation function for photonic neural networks, *Opt. Mat. Exp.* 8, 12, p. 3851, (2018).
[3] J. K. George, et al. Noise and Nonlinearity of Electro-optic Activation Functions in Neuromorphic Compute Systems, *Opt. Exp.* 27, 4 (2019).
[4] A. Mehrabian, et al. A Winograd-based Integrated Photonics Accelerator for Convolutional Neural Networks, IEEE J. Sel. Top. Qua. El. (2019).
[5] We acknowledge support from the Office of Navy Research under award number N00014-19-1-2595 of the Electronic Warfare program.