Advanced Photonics Congress (BGPP, IPR, Networks, NOMA, PVLED, SOF, SPPCom) © Optica Publishing Group 2022

Highly Integrated Photonic Tensor Core for imagining processing

Xiaoxuan Ma[1], Nicola Peserico[1], Ahmed Khaled[2], Zhimu Guo[2], Jiawei Meng[1], Bhavin J. Shastri[2], and Volker J. Sorger[1]

[1]*Department of Electrical and Computer Engineering, The George Washington University Washington, D.C., USA*

[2]*Department of Physics, Engineering Physics and Astronomy, Queen's University 99 University Ave, Kingston, ON K7L 3N6, Canada*

E-mail: sorger@gwu.edu

Abstract: Here we present a highly integrated architecture to perform Matrix-Vector Multiplication. By using the GEMM compiler, we can process images for 3-bits edge detection and 5-bits blur filter, with an error rate lower than 5.1%. © 2022 The Author(s)

1. Introduction

The explosion of data processing has pushed the common computing architectures toward their limits in terms of speed and throughput [1]. In particular, Neural Network algorithms are not well suited for common CPUs, as they rely much on the Multiplication and Accumulation (MAC) operation, that CPUs are not designed to perform efficiently [2]. To overcome this limitation novel architectures and paradigms have been developed [3-4]. Among many, optical computing has shown important progress thanks to the almost unlimited bandwidth, the high energy efficiency on performing MAC operations, and the high integration. In particular, in recent years many architectures have been proposed exploiting the electromagnetic nature of the light propagating in nanoscale structure to perform multiplication and accumulation [5-9]. However, till now, all those have missed the high integration and parallelism that Silicon Photonics can add to the circuits.

Here, we design and demonstrate a highly integrated Silico Photonic Tensor Core. By using two columns of micro-ring resonators with three different resonance frequencies, we achieve on-chip WDM fan-out and combiner. Our weight is built by the thermal-optical Mach-Zehnder modulator. The adder is built by the on-chip photodiode. Based on this PTC structure, image convolution is achieved, with an error rate of the 3 bits edge detection kernel is 1.1% and the 5 bits blur filter kernel is 5.05%.

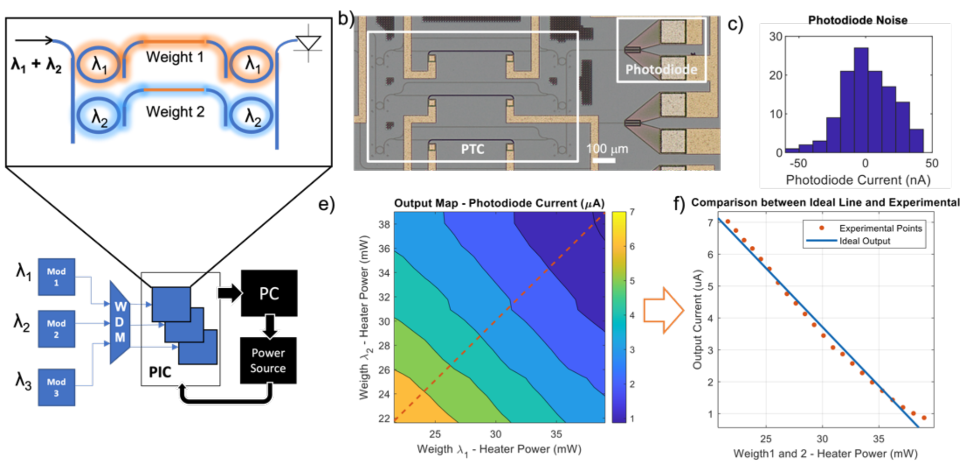

Fig. 1. a) Photonic tensor core architecture. b) Optical micrograph of photonic tensor core. c) Noise analyse of on-chip photodiode d) The schematic of testing setup for PTC e) Power map of one photonic tensor core. f) Performance of MAC operation on-chip photodiode

2. Results and Discussion

2.1. Results

For performing matrix-vector multiplication (MVM), the input vector is encoded in different wavelengths. Through the on-chip fan-out wavelength-division multiplexing (WDM), different wavelengths are separated and sent to the

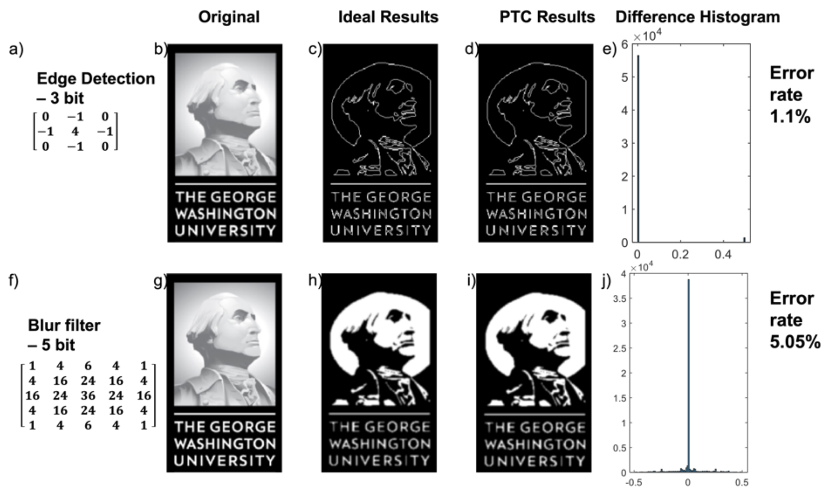

Figure 2: a) 3-bits Edge detection kernel. b-e) Results of edge detection. f) 5-bits Blur filter kernel. g-j) Results of blur filter

weight bank. The weighted light with different wavelengths is combined by the WDM combiner and added by the on-chip photodiodes (Fig. 1a).

We designed and taped out 3×3 photonic tensor core (PTC) from Advanced Micro Foundry (AMF). The single PTC includes a WDM Fan-out, a weight bank, a WDM combiner, and a photodiode (Fig. 1b). Each WDM includes three micro-ring resonators (MRRs) with different resonance wavelengths. The weight bank built by the thermal-optical Mach-Zehnder modulator array weighted the optical power of different wavelengths. To test this photonic tensor core, two lasers with different wavelengths are used. The off-chip modulators and WDM are used to encode and combine the light based on the input vector. The encoded laser is sent to the PIC to be weighted. The result of MVM is read by computer from the photodiodes. Also, the power source for setting the weight bank is controlled by the computer (Fig. 1d).

We initialize the PTC and select wavelengths by tuning the lasers to achieve an equal output power on each wavelength channel. Also, the noise of the photodiode is ±25 nA (Fig. 1c). For performing dot product, the power map with fixed input vector is tested. With different voltages applied to different MZMs, the output power in different wavelengths is modulated. The Light Power in different wavelengths is accumulated in photodiodes. The gradient change of total power is shown in the power map (Fig 1e-f). By using a power map as a look-up table (LUT), the edge detection based on our photonic tensor core is achieved. The 3 bits kernel [0 -1 0; -1 4 -1; 0 -1 0] is encoded by voltage and sent to weight bank. The result of MVM is selected from the power map. Figure b-e are the original picture and ideal result of edge detection, and the PTC result of edge detection. The difference between PTC result and the ideal result shows the error rate of our PTC is 1.1%. To exploir our PTC performance of higher bit resolution, 5-bits blur filter [1 4 6 4 1; 4 16 24 16 4; 16 24 36 24 16; 4 16 24 16 4; 1 4 6 4 1] is also tested. From the comparsion between figure h and figure i, 5.05% error rate is achieved.

2.2. Discussion

We have shown a photonic tensor core with on-chip WDM. By using the photonic tensor core, the 5 bits blur filter of the image is achieved. Future steps will include the integration of lasers as well as low VpiL modulators for a full packaging of the PTC.

3. References

[1] D. Amodei, “AI and Compute,” https://openai.com/blog/ai-and-compute (2020).
[2] Cong and B. Xiao, "Minimizing Computation in Convolutional Neural Networks", *Artificial Neural Networks and Machine Learning – ICANN 2014*, (2014), pp. 281-290.
[3] J. Peng et al., "DNNARA: A Deep Neural Network Accelerator using Residue Arithmetic and Integrated Photonics", *49th International Conference on Parallel Processing - ICPP*, (2020)
[4] Y. Shen et al., "Deep learning with coherent nanophotonic circuits", *Nature Photonics*, vol. 11, no. 7, pp. 441-446, (2017).
[5] N. Peserico et al., "Emerging devices and packaging strategies for electronic-photonic AI accelerators: opinion", *Opt. Mater. Express*, 12, 1347-1351 (2022)
[6] J. Feldmann et al., "Parallel convolutional processing using an integrated photonic tensor core", *Nature*, vol. 589, pp. 52-58, (2021).
[7] M. Miscuglio et al., "Photonic tensor cores for machine learning", *Applied Physics Reviews* 7, 031404 (2020).
[8] Shastri, B.J et al., "Photonics for artificial intelligence and neuromorphic computing", *Nat. Photonics* 15, 102–114 (2021).
[9] X. Ma et al., " Photonic Tensor Core With Photonic Compute-in-Memory ", *Optical Fiber Communication Conference,* (2022).