Efficient Photonic Deep Learning Leveraging 3D Optoelectronic Integration

Nathan Youngblood, Department of Electrical and Computer Engineering

Motivation

- Deep neural networks (DNNs) require massive computing resources for moderate improvements (e.g., ~500× more processing to achieve 2× improvement in accuracy [1]).
- Optical processors can operate at much higher speeds than electrical processors [2] but are limited in scalability and challenging to fabricate.

Project Description

- We aim to demonstrate a coherent photonic-electronic prototype which accelerates matrix operations for DNNs.
- Our prototype will combine a photonic integrated circuit with an off-the-shelf image sensor for scalable readout at quantum-limited efficiencies.

Context

- Our proposed approach addresses three major challenges hindering both photonic and electronic analog DNN accelerators:
  - Sensitivity to fabrication variability
  - High-speed electrical readout
  - Frequent reprogramming of analog weights

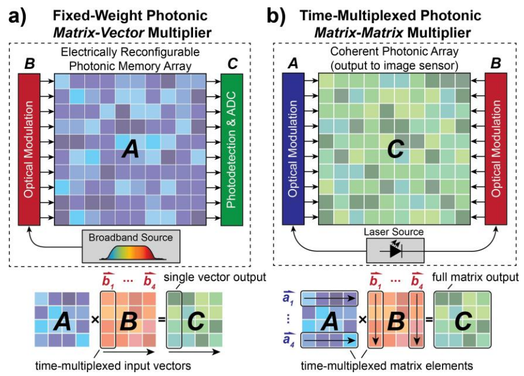

Figure 1: Comparison with other photonic architectures. Unlike fixed-weight photonic approaches **(a)** which are limited by the size of the memory array and can require a broadband light source, we propose a time-multiplexed architecture **(b)** where matrices ***A*** and ***B*** are encoded in the optical field. This decouples the optical modulation frequency from the speed of electrical readout, significantly improving compute efficiency.

University of Pittsburgh

We aim to accelerate deep learning applications by designing **fast**, **efficient**, and **scalable** optical processors.

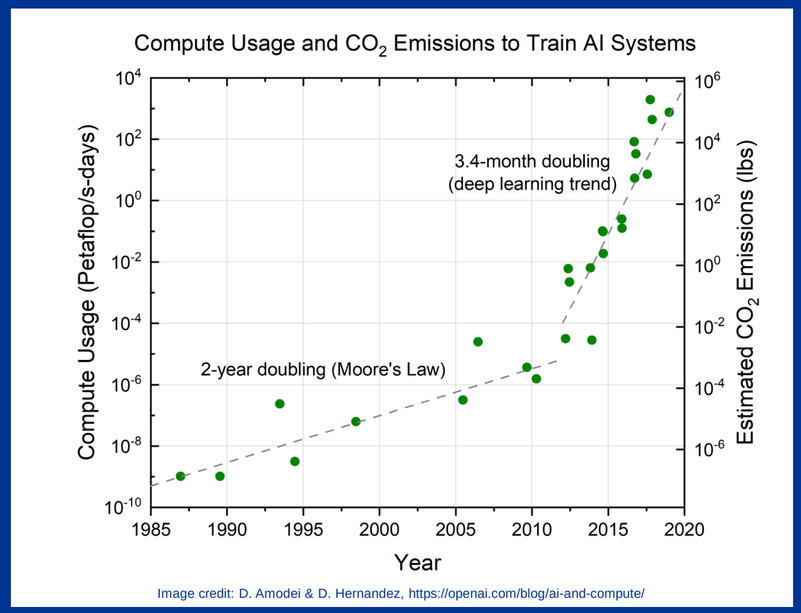

Image credit: D. Amodei & D. Hernandez, https://openai.com/blog/ai-and-compute/

Project Deliverables

- This work will result in the first experimental demonstration of a photonic matrix-matrix multiplier using a time-multiplexed approach.
- In the following year, we will leverage our platform to demonstrate high speed neuromorphic computing using time-dependent activation kernels.
- Project success will be determined by:
  - Fabrication and testing of integrated photonic components (mid-term goal)
  - Functional imaging system built for optoelectronic readout (mid-term goal)
  - Demonstration of coherent matrix-matrix multiplication (final goal)
  - Simulation model to benchmark efficiency and latency (final goal)
  - Submission of CAREER proposal, journal article, and conference paper (final goal)

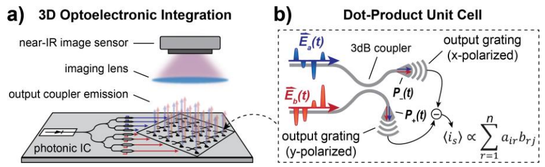

Figure 2: Schematic of proposed architecture. (**a**) A photonic integrated circuit with an array of dot-product unit cells performs multiply operations on-chip through coherent homodyne detection. An off-chip image sensor sums up the intensity and reads out the final matrix result. **(b)** Each unit cell contains a photoelectric multiplier to achieve the dot product between two time-multiplexed optical signals [3]. Light is coupled out-of-plane to the near-IR image sensor using grating couplers to take full advantage of both 2D and 3D integration.

Potential Impact

- We estimate our approach can improve computational efficiency by >100× (<10 fJ/OP) while decreasing inference latency by >10× compared to state-of-the-art approaches [4,5]. This would significantly reduce the carbon footprint of deep learning architectures based on digital electronics (e.g., GPUs).
- Using a fabrication-tolerant design, simple passive components, and off-the-shelf image sensors, we envision a quicker path to commercial readiness than competing approaches.

References and Acknowledgements

1N. C. Thompson et al., ***MIT Initiat. Digit. Econ. Res. Br.*** 4 (2020)

2 Feldmann* / N. Youngblood* / M. Karpov* et al., ***Nature*** 589, 52 (2021)

3 R. Hamerly et al., ***Phys. Rev. X*** 9, 021032 (2019)

4 N. P. Jouppi et al., ***Commun. ACM*** 63, 67 (2020)

5 Nvidia, "Nvidia V100 Tensor Core GPU," (2020) https://images.nvidia.com/content/technologies/volta/pdf/volta-v100-datasheet-update-us-1165301-r5.pdf

- Pitt Momentum funds are matched by the generous support of the ECE Department.