Towards Area-Efficient Optical Neural Networks: An FFT-based Architecture
Jiaqi Gu¹, Zheng Zhao¹, Chenghao Feng¹, Mingjie Liu¹, Ray T. Chen¹, David Z. Pan¹
¹University of Texas at Austin

Multi-layer Perceptron Inference
- Input: Vector $x$
- Output: Vector $y = \sigma(Wx)$
- Objective: Accuracy

Proposed ONN Architecture
- Optical Fast Fourier Transform
  - $X_k = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x_n e^{-jk\pi N n}$, $k = 0, 1, \ldots, N-1$
- 2 × 2 couplers and phase shifters to achieve OFFT

Element-wise Vector Multiplication
- Complex-valued multiplication
- Light polarization
- Encode weights in attenuators

Experimental Results
- Numerical Simulation (Lumerial)
- Normalized Area

Conclusion and Future Work
- New architecture to save optical component for better area efficiency
- Enable structured pruning to optical neural networks for network slimming without accuracy degradation
- 2.2–3.7x better area cost than SVD-based architecture
- May extend to CNN and other compact NNs
- Considering more practical hardware information

Hardware Utilization Analysis
- SVD-based Architecture ($W \in \mathbb{R}^{m \times n}$)
  - $\#DC_{SVD} = m(n-1) + n(n-1) + \max(m, n)$
  - $\#PS_{SVD} = \frac{m(m+1)}{2} + \frac{n(n+1)}{2}$
- TΣU-based Architecture ($W \in \mathbb{R}^{m \times n}$)
  - $\#DC_{TΣU} = n(n+1) + \max(m, n)$
  - $\#PS_{TΣU} = \frac{\min(n, m)}{2}(2 \log_k k + 1)$

Source code: https://github.com/JeremieLeco/fft-onn