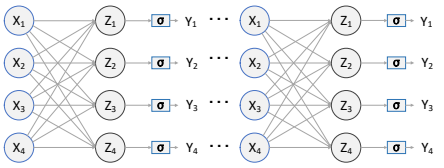
FLOPS: EFficient On-Chip Learning for OPtical Neural Networks Through Stochastic Zeroth-Order Optimization

Jiaqi Gu[1], Zheng Zhao[1], Chenghao Feng[1], Wuxi Li[2], Ray T. Chen[1], David Z. Pan[1]

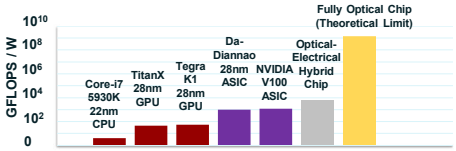
[1]University of Texas at Austin, [2]Xilinx Inc.

Neural Networks and AI Acceleration

- Input
  - Vector x
- Output
  - Vector $y = \sigma(W \cdot x)$

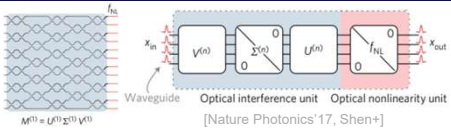


- ML Applications and Photonic Acceleration



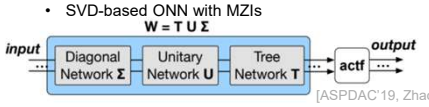
[Nature Photonics'17, Shen+]

ONN Architectures



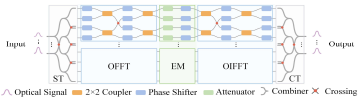
[Nature Photonics'17, Shen+]

- SVD-based ONN with MZIs



[ASPDAC'19, Zhao+]

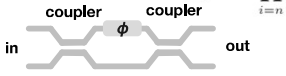
- $T\Sigma U$-based ONN with MZIs and sparse tree

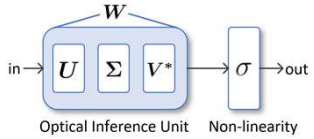


- FFT-based area-efficient ONN [ASPDAC'20, Gu+]

Principles of MZI-based ONNs

- **Singular value decomposition (SVD)** $W = U\Sigma V^*$
- **Unitary group parametrization** $U(n) = D \prod_{i=n}^{2} \prod_{j=1}^{i-1} R_{ij}$

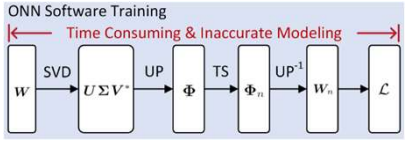




Previous ONN Training Protocols: Software-based

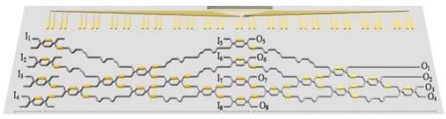
- **Software training**
- **Unaware of optical hardware**
- **Limited speed (>1 s per iteration)**

$$\Phi^* = \arg\min_{\Phi} L(\Phi)$$



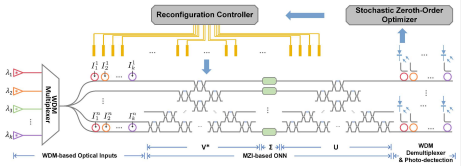
Previous ONN Training Protocols: On-chip

- **On-chip ONN training**
- **2-3 order-of-magnitude faster than software training**
- **~ 1ms per iteration**
- **Unscalable: ~100 MZIs**
- **Limited efficiency: evolution-based search**



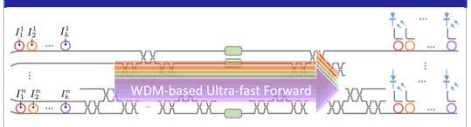
[arXiv 2019, Zhou+]

Proposed Method: FLOPS



- **On-chip learning via stochastic zeroth-order optimization**
- **Efficiency**
  - WDM-based parallel gradient estimation
- **Accuracy**
  - Two-stage learning protocol with high accuracy
- **Robustness**
  - Robust learning under *in situ* device variations

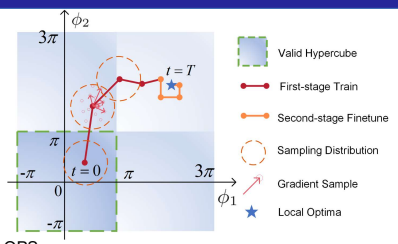
Zeroth-order Gradient Estimation



- **Higher parallelism**
  WDM-based asymmetric gradient estimator
- **Higher efficiency**
  Numerical differentiation without back-propagation
- **Lower variance**
  Multiple gradient samples
- **Faster convergence**
  First-order gradient descent

$$\nabla_\Phi L = \frac{1}{Q\sigma^2} \sum_{q=0}^{Q-1} \left(L_S(x; \Phi + \Delta\Phi_q) - L_S(x; \Phi) \right) \Delta\Phi_q$$

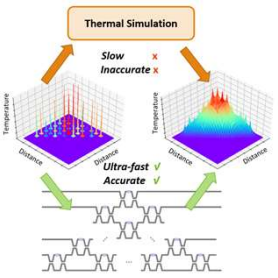
Extended FLOPS+

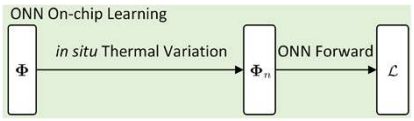


- FLOPS
  - Fast exploration in phase space
  - Sample-efficient
  - Not accurate enough
- FLOPS+ with ***SparseTune***
  - Sparse coordinate-wise fine-tuning
  - Improve accuracy via searching
  - Sparsity guarantees efficiency

Robust Learning under Thermal Variations

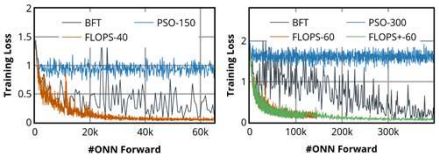
- **Thermal Crosstalk**
  - Time-consuming
  - Inaccurate
- **On-chip handling**
  - Ultra-fast: ~1 μs
  - Accurate modeling





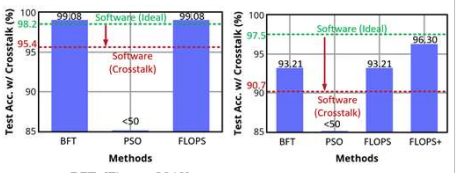
Experimental Results

- **Efficient learning on vowel recognition task**
  - 2-4x more query efficient than BFT and PSO



ONN config: 8-16-16-4 (448 MZIs) ONN config: 10-24-24-6 (960 MZIs)

- **Robust learning under *in situ* thermal variation**
  - 3-5% higher accuracy than previous methods



BFT: [Zhou+, 2019] PSO: [Zhang+, 2019]

Contribution

Efficient on-chip learning for MZI-based ONNs

- **$O(10^3)$ more efficient than software training**
- **3% higher accuracy by FLOPS+**
- **2-4x faster learning than prior on-chip training algorithms**
- **More scalable on ~1000 MZIs (prior work ~100 MZIs)**
- **5% higher accuracy under thermal variations**

Conclusion and Future Work

- Extend FLOPS to more advanced ONN architecture
- Further develop better algorithms for efficiency improvement
- Online learning and unsupervised learning on photonic chips
- Photonic chip tape-out and testing