Hardware-Efficient Photonic Tensor Cores Based on Differential Multi-Operand Mach-Zehnder Interferometers

Zhenxiang Xu,[1,†] Shupeng Ning,[1,†] Chenghao Feng,[1] Hanqing Zhu,[1] David Z. Pan,[1] Jiaqi Gu,[1,2] and Ray T. Chen[1,*]

[1] *Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78758, USA*
[2] *School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ 85281, USA*
[†] *These authors contributed equally to this work.*

[*] *chenrt@austin.utexas.edu*

Abstract: We propose a compact architecture using multi-operand Mach-Zehnder interferometers (MOMZIs) for full-range photonic tensor operations with high hardware efficiency and computational density. Experiments show low inference error and strong performance on representative machine-learning tasks.

1. Introduction

Deep neural networks (DNNs) have driven major advances in artificial intelligence (AI), but their rapidly increasing computational demands strain conventional electronic accelerators. Bandwidth and power constraints in electronics have motivated alternative computing paradigms such as optical computing based on photonic integrated circuits (PICs), which exploit the intrinsic advantages of light for energy-efficient multiply-accumulate (MAC) operations [1, 2]. However, existing photonic tensor cores (PTCs) still suffer from limited scalability due to large device footprints, accumulated optical loss, and high control complexity [3, 4].

To overcome these bottlenecks, we develop a hardware-efficient PTC based on MOMZIs. The proposed differential unit supports full-range tensor encoding operations within a compact footprint. Experimental results demonstrate accurate on-chip inference and high accuracy on image classification tasks, highlighting the stability and precision of the demonstrated optical computing scheme.

2. Design and Working Mechanism

2.1. *Multi-operand Mach-Zehnder Interferometers*

Compared with conventional Mach-Zehnder modulators, the proposed MOMZI integrates k independent phase shifters (PSs) in each arm, allowing simultaneous modulation by multiple signals. As shown in Fig. 1(a)&(b) , each differential unit consists of a pair of MOMZIs and balanced photodetectors (PDs). In each MOMZI, one arm is applied by $(w_i + x_i)$ or $(w_i - x_i)$, while the other arm provides a bias that sets the initial state at the onset of the most linear region of its sinusoidal transfer function (Fig. 1(c)). Owing to the quadratic dependence of the phase shift on the drive signal in thermo-optic or Kerr-effect modulators, the output photocurrent scales as $\sum_i (w_i \pm x_i)^2$. By subtracting the two photocurrents, the quadratic terms cancel, yielding the inner product $\boldsymbol{w} \cdot \boldsymbol{x}$. It should be noted that MOMZIs can encode two *full-range dynamic* operands on a single actuator (e.g., the two terminals of a PS), which avoids using multiple modulators to separately encode x_i and w_i as in conventional architectures [2, 4]. Furthermore, extensive sharing of drive signals, including both between MOMZI pairs and within each unit for weight sharing, leads to a compact, layout-friendly PIC topology with simplified electrical routing.

2.2. *Architecture of Optical Neural Networks Based on MOMZI Differential Units*

We propose an optical neural network (ONN) architecture based on MOMZIs. As illustrated in Fig. 1(a), an $M \times N$ weight matrix is partitioned along each row into Q segments of length k. Each MOMZI pair performs a length-k dot product within a single differential unit, enabling dense matrix-vector multiplication (MVM) through spatial multiplexing and temporal integration. Since MOMZIs support dynamic encoding of both operands, the architecture is also compatible with *Transformer* attention blocks [5], which require multiplications between two dynamic matrices rather than fixed weights.

3. Results and Discussion

In this work, we taped out a photonic-electronic neural chip with thermo-optic MOMZIs, each unit integrating four operands. For characterization, we fitted the transmission characteristics of each MOMZI using random length-4 input vectors and then biased it in the linear operating region (Fig. 1(c)).

To further evaluate the functionality of MOMZIs, we implemented an on-chip image processing task using multiple 3×3 convolution kernels applied to an image of the UT Tower (Fig. 2(a)). As shown in Fig. 2(b), the results exhibit the expected features with high fidelity to the digital reference. Additionally, we trained and

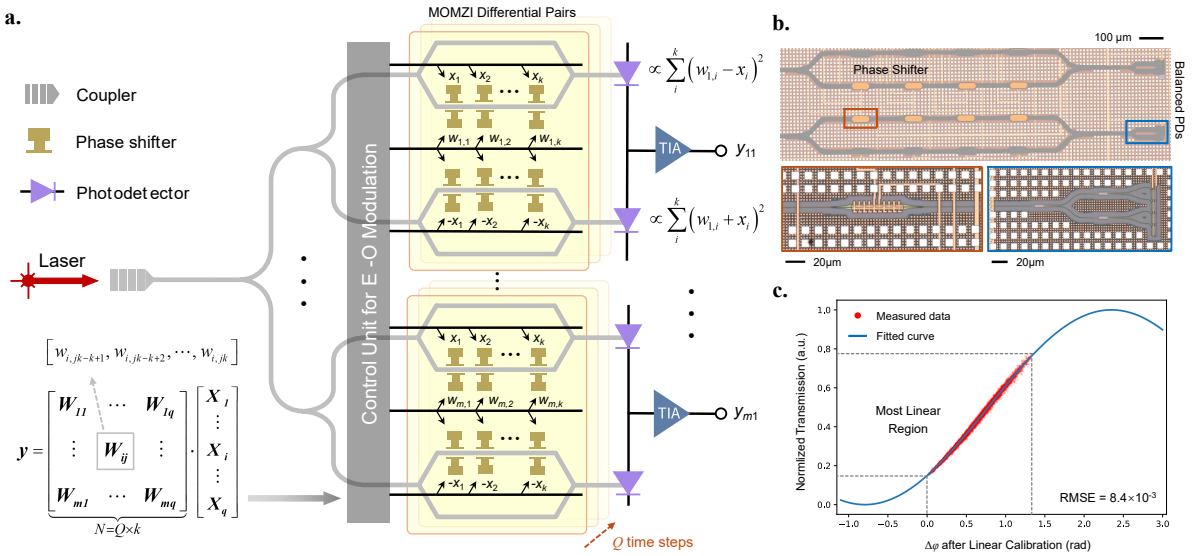

Fig. 1: (a) Schematic of the MOMZI and the proposed PTC architecture. (b) Micrographs of a fabricated MOMZI. (c) Measured MOMZI transmission response and fitted curve, with the linear operating region highlighted.

deployed a 2-layer convolutional neural network (CNN) for image classification on the Fashion-MNIST dataset, achieving 95.0% training accuracy and 91.5% experimentally measured test accuracy (Fig. 2(c)).

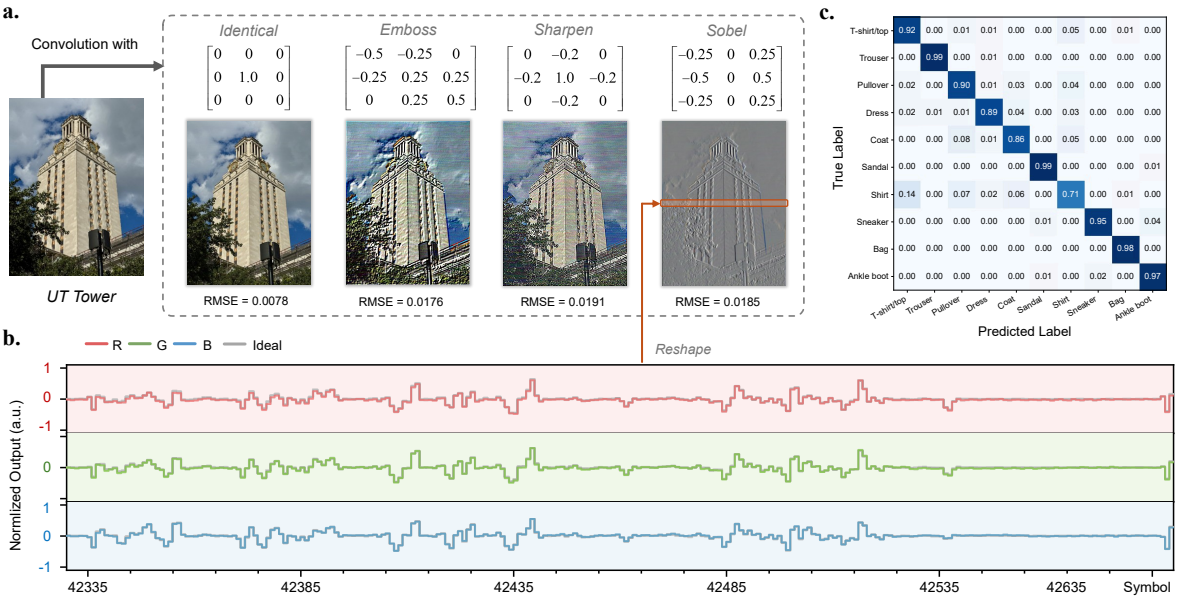

Fig. 2: Experimental results of on-chip image processing and classification tasks. (a) UT Tower image and the feature maps extracted by different kernels. (b) Measured and ideal waveforms of the convolution outputs. (c) Confusion matrix for Fashion-MNIST image classification with on-chip inference.

4. Conclusion

In this work, we demonstrated a MOMZI-based PTC with high compactness and hardware efficiency. Experimental results show small deviations in on-chip tensor operations and DNN inference. The proposed PTC is compatible with a variety of modern AI architectures and helps alleviate the scalability bottleneck of ONNs. This work is supported by Air Force Office of Scientific Research and the Multidisciplinary University Research Initiative.

References

1. Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature **521**(7553), 436–444 (2015).
2. Y. Shen, N. C. Harris, S. Skirlo, M. Prabhu, T. Baehr-Jones, M. Hochberg, X. Sun, S. Zhao, H. Larochelle, D. Englund, and M. Soljačić, “Deep learning with coherent nanophotonic circuits,” Nature Photonics **11**(7), 441–446 (2017).
3. S. Ning, H. Zhu, C. Feng, J. Gu, D. Z. Pan, and R. T. Chen, “Hardware-efficient photonic tensor core: Accelerating deep neural networks with structured compression,” Optica **12**(7), 1079–1089 (2025).
4. A. N. Tait, T. F. De Lima, E. Zhou, A. X. Wu, M. A. Nahmias, B. J. Shastri, and P. R. Prucnal, “Neuromorphic photonic networks using silicon photonic weight banks,” Scientific Reports **7**(1), 7430 (2017).
5. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in neural information processing systems **30** (2017).