Evaluation of a compact butterfly-style photonic-electronic neural chip on complicated deep learning tasks

Chenghao Feng [1], Jiaqi Gu [2], Hanqing Zhu [2], Rongxing Tang [1], David Z. Pan [2], and Ray T. Chen [1*]
[1]*Microelectronics Research Center, The University of Texas at Austin, Austin, Texas 78758, USA*
[2]*Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, Texas 78705, USA*
[*] *e-mail address: chenrt@austin.utexas.edu*

Abstract: We deploy a compact butterfly-style photonic-electronic neural chip on ResNet-20 and achieve > 85% measured accuracy on the CIFAR-10 dataset with only 3-bit weight programming resolutions, showing its practicality in implementing complicated deep learning tasks. © 2023 The Author(s)
OCIS codes: (200.4700) Optical neural systems; (200.4260) Neural networks; (130.3120) Integrated optics devices

1. Introduction

The optical neural network (ONN) is a promising analog artificial intelligence (AI) accelerator that utilizes some unique features of light, such as low latency, low power consumption, and high parallelism. Earlier work has presented various high-performance integrated ONNs to implement general matrix multiplications (GEMMs) in deep neural network (DNN) models [1–3]. However, GEMM-based ONNs require unnecessarily large area cost and high control complexity. In our previous work [4], we devise an optical subspace neural network architecture (OSNN) (Fig. 1(a)) based on butterfly-style photonic meshes, which trades the universality of weight representation for lower optical component usage, area cost, and energy consumption. We experimentally demonstrate the utility of OSNN on our butterfly-style photonic neural chip (BPNC) on the MNIST dataset with a measured accuracy of 94.16%. After proving the practicality of OSNN on simple hand-written digit recognition tasks, it is interesting to investigate whether our OSNN with restricted matrix parameter space has enough model expressivity to implement more complicated deep learning tasks.

In this study, we deploy our 4 × 4 BPNC (shown in Fig. 1(b)) in large DNN models, e.g., ResNet-20 with 0.27 million parameters, to evaluate the task performance of OSNN on more complicated datasets such as CIFAR-10. Our initial training results show that our BPNC achieves a measured accuracy of 85.12% under only a 3-bit weight programming resolution on the CIFAR-10 dataset, which is comparable to the accuracy of a 64-bit computer. This work proves that our hardware-efficient OSNN has enough expressivity to implement complicated deep learning tasks in the future.

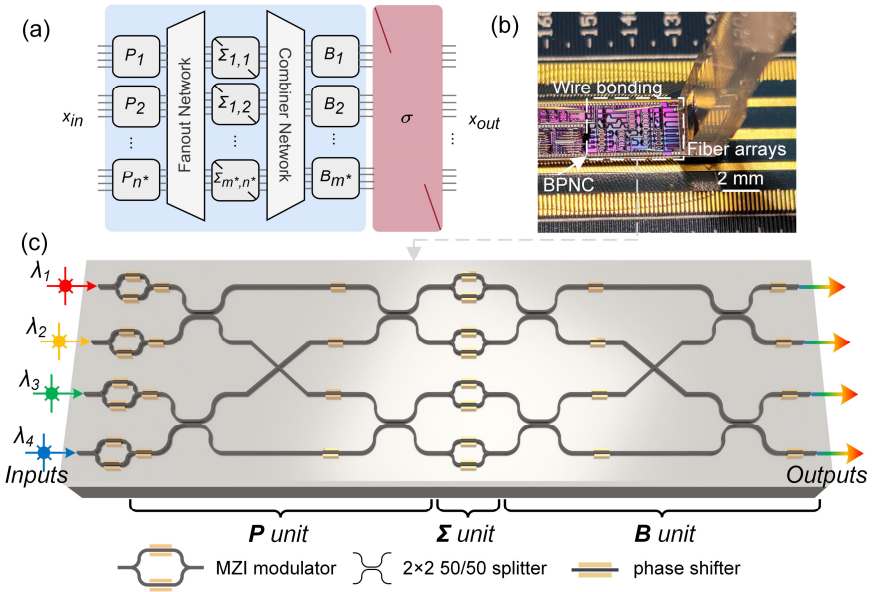

Fig. 1. Schematic of the optical subspace neural network (OSNN) architecture. (a) shows a n-input, m-output layer which consists of $n^* = \frac{n}{k}$ projection units (***P*** units), $m^* = \frac{m}{k}$ butterfly-style transform units (***B*** units), $m^* \times n^*$ diagonal matrix units (***Σ*** units), and an electrical $\boldsymbol{\sigma}$ unit to implement activation functions. (b) shows the photograph of our packaged 4 × 4 butterfly-style photonic-electronic neural chip (BPNC). (c) shows the schematic of the 4 × 4 BPNC under the multi-wavelength-input mode, where different photonic circuit units (***P*/*Σ*/*B***) are shown.

2. Optical subspace neural network architecture

The schematic of the OSNN architecture is shown in Fig. 1, where an $m \times n$ weight matrix in one fully-connected layer is partitioned into $\frac{mn}{k}$ $k \times k$ ($k = 4$ or 8) submatrix units. Shown in Fig. 1(c), each sub-matrix unit is composed of two $k \times k$ butterfly-style photonic meshes ($\boldsymbol{P}$ and $\boldsymbol{B}$ units) and a diagonal matrix unit ($\boldsymbol{\Sigma}$ unit) consisting of a column of k modulators. Each $\boldsymbol{B}/\boldsymbol{P}$ unit is shared by multiple $\boldsymbol{\Sigma}$ units, and only $\frac{n}{k}$ P units, $\frac{m}{k}$ B units are required to implement an n-input, m-output layer, resulting in a smaller chip footprint than in previous work [5]. Moreover, only the $\frac{mn}{k}$ active photonic devices in the Σ unit are trainable. The total number of trainable components is then $k - 1$ times smaller than ONNs designed for GEMMs [2], significantly reducing the weight loading cost and reprogramming complexity.

Our OSNN architecture supports both single-wavelength and multi-wavelength optical inputs, and each 4×4 $\boldsymbol{B\Sigma P}$ block can realize 64.4% fidelity in expressing general 4×4 matrices in single-wavelength mode and 92.2% fidelity in expressing arbitrary all-positive or all-negative 4×4 matrices. Details of our OSNN architecture are disclosed in [4].

3. Results

In this work, we deploy our BPNC on ResNet-20 to implement image classification tasks. Large tensor operations in ResNet-20 are partitioned into 4×4 blocks and mapped onto our BPNC. In the testing, the input lights with different wavelengths (1548, 1549, 1550, 1551 nm) are generated by a tunable laser. A multi-channel DAC is required to program the input and weight signals of the BPNC. We then use two 2-channel oscilloscopes to read the output signals. The measurement data is modeled and processed using our hardware-aware on-chip training framework [4]. Finally, we offload other functions, such as nonlinear activation and partial accumulations in computers to implement the DNN model.

The testing results are shown in Fig. 2, which shows that we can achieve a measured accuracy of 85.12% under a 3-bit weight control resolution, which is close to the accuracy of a 64-bit computer (~90%). We will further improve the task performance by optimizing our training framework. We will show more benchmarking results on different DNN models and complicated datasets in the presentation.

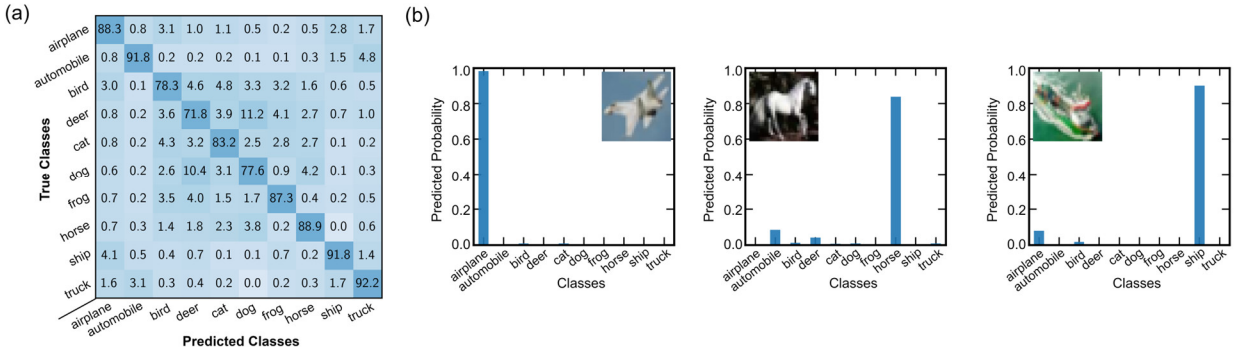

Fig. 2. Testing results of 4×4 BPNC using ResNet-20 model on the CIFAR-10 dataset. The BPNC is working under multi-wavelength mode with a 3-bit weight control resolution. (a) The confusion matrix of our benchmarking results shows a measured accuracy of 85.12%. (b) shows some output probability distribution of three images of different classes.

In conclusion, we deploy our 4×4 BPNC on ResNet and evaluate its performance on the CIFAR-10 dataset. Our OSNN architecture and hardware-aware DNN deployment approach provide a synergistic solution to implement complicated deep learning tasks with similar task performance but with lower component usage, smaller footprint, and lower power consumption than ONNs architectures designed for GEMMs.

References

1. Y. Shen et al., "Deep learning with coherent nanophotonic circuits," Nature Photonics **11**, 441–446 (2017).
2. A. N. Tait et al., "Neuromorphic photonic networks using silicon photonic weight banks," Scientific Reports **7**, 7430 (2017).
3. A. Sludds et al., "Delocalized photonic deep learning on the internet's edge," Science **378**, 270–276 (2022).
4. C. Feng et al., "A compact butterfly-style silicon photonic-electronic neural chip for hardware-efficient deep learning," arXiv preprint arXiv:2111.06705 (accepted by ACS photonics) (2022).
5. J. Gu et al., "Toward Hardware-Efficient Optical Neural Networks: Beyond FFT Architecture via Joint Learnability," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems **40**, 1796–1809 (2021).

The authors acknowledge support from the Multidisciplinary University Research Initiative (MURI) program through the Air Force Office of Scientific Research (AFOSR) (Grant No. FA 955017-1-0071) monitored by Dr. Gernot S. Pomrenke.