POLITECNICO DI TORINO
Repository ISTITUZIONALE

Analysis of the Hardware Imprecisions for Scalable and Compact Photonic Tensorized Neural Networks

Original
Analysis of the Hardware Imprecisions for Scalable and Compact Photonic Tensorized Neural Networks / On, M. B.; Lee, Y. -J.; Xiao, X.; Proietti, R.; Ben Yoo, S. J.. - ELETTRONICO. - (2021), pp. 1-4. (Intervento presentato al convegno 2021 European Conference on Optical Communication, ECOC 2021 tenutosi a Bordeaux, France nel 13-16 September 2021) [10.1109/ECOC52684.2021.9605948].

Availability:
This version is available at: 11583/2973263 since: 2022-11-22T08:53:00Z

Publisher:
Institute of Electrical and Electronics Engineers Inc.

Published
DOI:10.1109/ECOC52684.2021.9605948

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
IEEE postprint/Author's Accepted Manuscript

©2021 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collecting works, for resale or lists, or reuse of any copyrighted component of this work in other works.

(Article begins on next page)

12 October 2024

Analysis of the Hardware Imprecisions for Scalable and Compact Photonic Tensorized Neural Networks

Mehmet Berkay On(1), Yun-Jhu Lee(1), Xian Xiao(1), Roberto Proietti(1), and S.J. Ben Yoo(1)

(1) Department of Electrical and Computer Engineering, University of California, Davis, One Shield Ave., Davis, California 95616 USA
sbyoo@ucdavis.edu

Abstract *We simulated tensor-train decomposed neural networks realized by Mach-Zehnder interferometer-based scalable photonic neuromorphic devices. The simulation results demonstrate that under practical hardware imprecisions, the TT-decomposed neural networks can achieve >90% test accuracy with 33.6× fewer MZIs than conventional photonic neural network implementations.*

Introduction

Photonic neural networks (PNNs) have demonstrated significantly improved energy efficiency and throughput over electronic artificial neural networks (ANNs) [1]. However, photonic neural networks' weight matrices are typically realized by Mach-Zehnder interferometer (MZI)-based [2] or by wavelength-division multiplexing (WDM)-based [3] architectures suffering from scalability challenges due to the required number of components. Tensor-train (TT) decomposition [4] is a promising method to overcome the challenges of dimensionality, especially for wide and deep neural networks while accompanying acceptably small amounts of performance degradations [5]. Through algorithm-hardware codesign, photonic tensorized TT-decomposed neural networks (TNN) can significantly improve scalability using a limited number of photonic components (i.e., MZIs). Our recent work demonstrated that TNNs could achieve a radix of N = 1024 using *1164*× fewer number of MZIs [6]. The decreased number of MZIs for the equivalent ANN implementations leads to lower optical losses, fewer electrical controls, and smaller die sizes for the desired radix. In this work, we investigated the effects of possible hardware imprecisions for photonic MZI-based TNNs. We simulated and compared the proposed architecture, a conventional photonic MZI-based ANN, and a 2-dimensional Fourier Transform (2D-FT) preprocessed ANN [7] (another method to scale PNNs). The hardware imprecisions such as phase-shifter variations and beam-splitter power imbalances are modeled in neural network simulations which perform MNIST handwritten digit classification. Our simulation results show that TNNs can outperform their counterparts in terms of accuracy besides reducing the number of required MZIs.

Photonic Tensorized Neural Networks

Fig. 1 visualizes the relationship between ANN's synaptic connections and TNN's synaptic connections. Synaptic interconnections of the multi-layer ANNs are represented as a *2*-dimensional weight matrix, $\boldsymbol{W}^t_{M\times N}$ where *M* is the number of neurons in *t-th* layer, *N* is the number of neurons in the *(t+1)-th* layer. Synaptic interconnections of the multi-layer TNNs are *d*-dimensional tensors $\boldsymbol{W}^t_{\mu_1\times v_1 \ldots \times \mu_d \times v_d}$ where $M = \prod_{i=1}^{d} \mu_i$, and $N = \prod_{i=1}^{d} v_i$. TT-decomposition can be interpreted as singular value decomposition (SVD) of multi-dimensional arrays [4]. After the decomposition, the tensor represented as, $\boldsymbol{W}^t(\mu_1(i), v_1(j), \ldots \mu_d(i), v_d(j)) = \prod_{k=1}^{d} \boldsymbol{G_d}(:, \mu_k(i), v_k(j), :)$ where $\boldsymbol{G_d}$'s are tensor cores with the shape of $R_{k-1} \times \mu_k \times v_k \times R_k$. R_k's are the rank of SVD. $R_0 = R_d = 1$ is defined as the boundary condition in TT-decomposition.

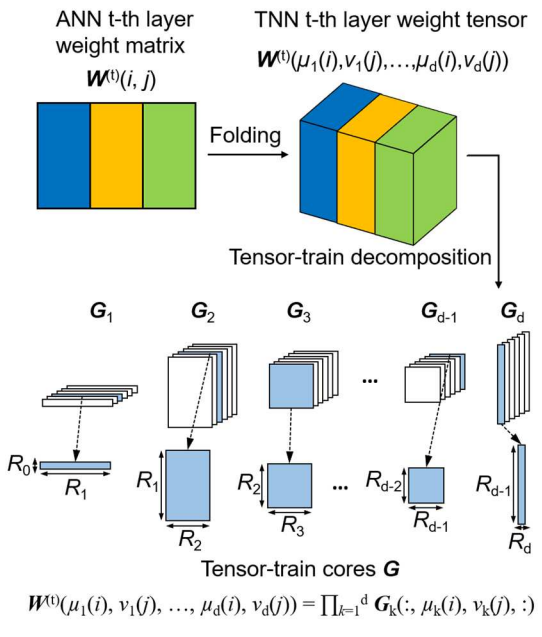

Fig. 1: Tensorized and TT-decomposed neural network weight matrix

978-1-6654-3868-1/21/$31.00 ©2021 IEEE

Then, the required matrix-vector multiplication operation in the ANNs, $\boldsymbol{y}^t_{1\times M} = \boldsymbol{W}^t_{M\times N}\boldsymbol{x}^t_{N\times 1}$, where $\boldsymbol{x}^t_{N\times 1}$ is the output of *t-th* layer's neurons, can be performed in the TNNs as:

$$\boldsymbol{y}^t_{\mu_1\times\ldots\mu_d}(i_1,\ldots i_d) = \sum_{j_1,\ldots j_d} \boldsymbol{g_1}\ldots\boldsymbol{g_d}\boldsymbol{x}^t_{v_1\times\ldots v_d}(j_1\ldots j_d)$$

where $\boldsymbol{g_k} = \boldsymbol{G_k}(:,\mu_k(i),v_k(j),:)$ are $R_{k-1}\times R_k$ matrices, $\boldsymbol{x}^t_{v_1\times\ldots v_d}$ is the tensorized output vector of the *t-th* layer's neurons. This computation can be realized as cascaded matrix-by-matrix products [4] by reshaping tensor cores. Then the entire TNN synaptic calculations can be realized by the MZI-based optical linear units (OLU).

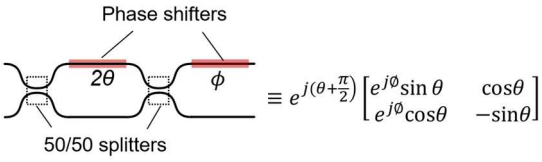

Fig. 2: Mach-Zehnder Interferometer and transfer matrix with perfect components

Fig. 2 shows the building block of the OLU and its transfer matrix, which consists of two 50/50 beam splitters and two phase-shifters. It is possible to construct an *N×N* arbitrary unitary matrix by using *N(N-1)/2* MZIs in either triangular [8] or rectangular [9] fashion. In this work we considered rectangular structures due to their compactness. SVD can be utilized by two arbitrary unitary matrices and *N* number of additional phase-shifters and attenuators to realize any arbitrary matrix. One of the impairments in the MZI is the variations in the phase-shifters θ, and $\emptyset$, which are modeled as normal gaussian noise with a standard deviation of σ in the neural network simulations. Another impairment is the beam splitters' power imbalance, which is modeled as the transfer function of the beam-splitter $\boldsymbol{T}_{BS} = \frac{1}{\sqrt{2}}\begin{bmatrix} a & jb \\ jb & a \end{bmatrix}$, where $a = b = 1$ in an ideal case, and $a^2 + b^2 = 2$. In the neural network simulations, the variations in the beam-splitters are modeled as $a^2 \sim N(1,\rho^2)$. Currently, there is no comprehensive study on the effects of these imperfections for TNNs.

Neural Network Simulation with Hardware Imprecisions

To evaluate the performance of the photonic TNN and compare it with a conventional photonic ANN and a 2D-FT preprocessed photonic ANN [7],we performed the MNIST handwritten digit classification task by using TensorFlow and t3f [10] python libraries as simulation platforms. The backpropagation algorithm trains each NN with the Adam optimizer in 10 epochs. Every neuron has rectified linear unit activation function and the categorical cross entropy loss evaluates the networks performance. We considered two types of ANN in simulations and TNNs as equivalent to them. The first ANN (*case (a)*) has single hidden layer with 128 neurons, while second (*case (b)*) has three hidden layers with 256 ,128, and 64 neurons, respectively. Before flattening the 28×28 grayscale MNIST inputs, images are cropped to 16×16 for *case (a)* and 20×20 for *case (b)* ANNs' input layers. Fig. 3 summarizes the implementation of a photonic ANN with OLUs, modulators, detectors, and hidden neuron circuits. At the output layer, the decision results can be observed directly after the detectors.

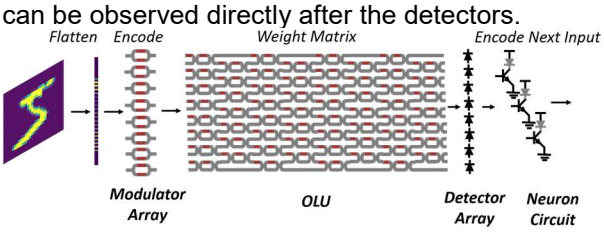

Fig. 3: Conventional Photonic ANN

The TNN equivalents of the *case (a)* tensorize the weight matrices, $W^1_{128\times256}$ and $W^2_{10\times128}$ as *4*-dimensional tensors, $W^1_{4\times4\times4\times4\times4\times4\times4\times2}$ and $W^1_{2\times5\times4\times2\times4\times1\times4\times1}$. Two different sized TNNs use TT-rank of 8 (*'TNN rank=8'* in Fig. 6(a) and Fig. 7(a)) and TT- rank of 4 (*'TNN rank=4'* in Fig. 6(a) and Fig. 7(a)) in simulations. In *case (b)*, TNNs tensorize the ANN weight matrices, $W^1_{256\times400}$, $W^2_{128\times256}$, $W^3_{64\times128}$, and $W^4_{10\times64}$ as *4*-dimensional tensors, $W^1_{4\times4\times5\times4\times5\times4\times4\times4}$, $W^2_{4\times4\times4\times4\times4\times4\times4\times2}$, $W^3_{4\times2\times4\times4\times4\times4\times2\times2}$, and $W^4_{2\times5\times4\times2\times4\times1\times2\times1}$. Three different sized TNNs are built. First one uses TT-rank of 8 for all tensors (*'TNN rank=8'* in Fig. 6(b) and Fig. 7(b)). Second one uses TT-rank of 8 for the tensors $\boldsymbol{W}^2$, $\boldsymbol{W}^3$, and $\boldsymbol{W}^4$, and TT-rank of 6 for the first tensor $\boldsymbol{W}^1$ (*'TNN rank=8,6'* in Fig. 6(b) and Fig. 7(b)). Lastly, third TNN uses TT-rank of 6 for the tensors $\boldsymbol{W}^2$, $\boldsymbol{W}^3$, and $\boldsymbol{W}^4$, and TT-rank of 4 for the first tensor $\boldsymbol{W}^1$ (*'TNN rank=6,4'* in Fig. 6(b) and Fig. 7(b)). Fig. 4 summarizes the photonic TNN implementation. Reindexing between the tensor cores can be realized by waveguide crossings, and the input layer vector can be tensorized with WDM modulators.

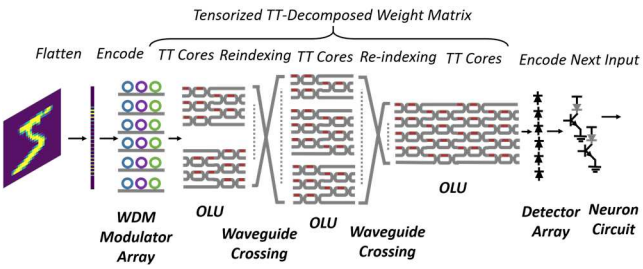

Fig. 4: Photonic TNN

Lastly, we build two 2D-FT preprocessed ANNs: *case (a)* by a single hidden layer with 32 neurons; and *case (b)* by two hidden layers with 64 and 32 neurons, respectively. We only used the amplitude of 70 complex-valued 2D-FT coefficients for *case (a)* and the amplitude of 100 complex-valued 2D-FT coefficients for *case (b)* around the central frequency. The sizes of the 2D-FT ANNs are chosen to match the number of required MZIs with '*TNN rank=8*' implementations. Fig. 5 summarized a possible 2D-FT preprocessed photonic ANN. 2D-FT can be performed by an off-chip 2f-optical system or by silicon photonic star couplers [11].

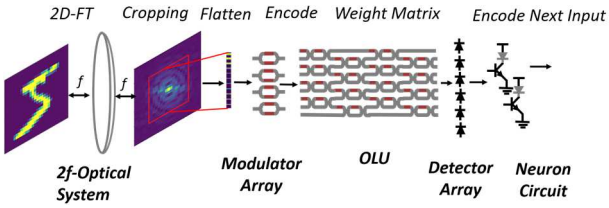

Fig. 5: 2D-FT Preprocessed Photonic ANN

Results and Discussion

We generated 50 different trials for each σ and ρ^2, and reported the mean and standard deviation of the test accuracies in the plots. Fig. 6 presents the effects of phase-shifter variations for the constructed neural networks. The robustness of the TNNs can be observed more clearly in *case (b)*. At $\sigma = 0.02$ radian, TNNs can achieve test accuracies >90% for both cases while saving up to *33.6*× MZIs resources.

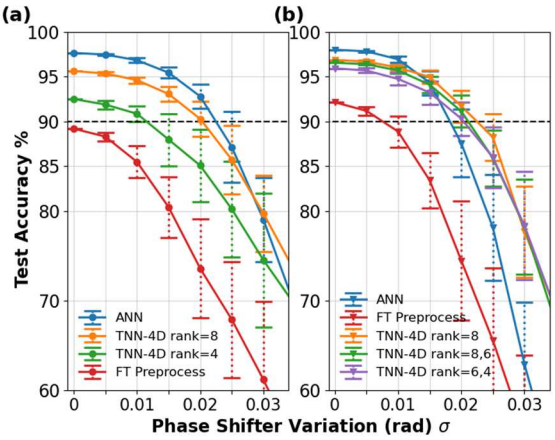

Fig. 6: Test Accuracy Percentage vs. Phase-Shifter Variation at $\boldsymbol{\rho^2 = 0}$, **(a)** *case (a)* ANNs, and equivalent TNNs**(b)** *case (b)* ANNs and equivalent TNNs

Fig. 7 shows the tolerance of the benchmarked neural networks against the beam-splitter power imprecisions at phase-shifter standard deviation $\sigma = 0.01$ rad. The results show that beam-splitter imperfections are more critical compared to phase-shifter variations. TNNs can still achieve >90% accuracy at $\sigma = 0.01\ rad$ and $\rho^2 = 0.25\ dB$ while saving up to *33.6*× MZIs resources. Although approximately the same amount of MZIs implements 2D-FT preprocessed ANNs, their test accuracies ~15% below the TNNs. The reason behind these observations is the increased OLU size. One can expect that for larger OLUs, the imperfections on the individual components will be averaged. The large OLUs are supposed to be more resilient than the compact, small-size OLUs. However, unlike for the WDM Microring weight banks implementation [3], the individual components inside the MZI-based OLU can affect multiple realized photonic weight matrices or tensors entries. These phenomena have been studied in [12], [13], where the authors reach a similar conclusion.

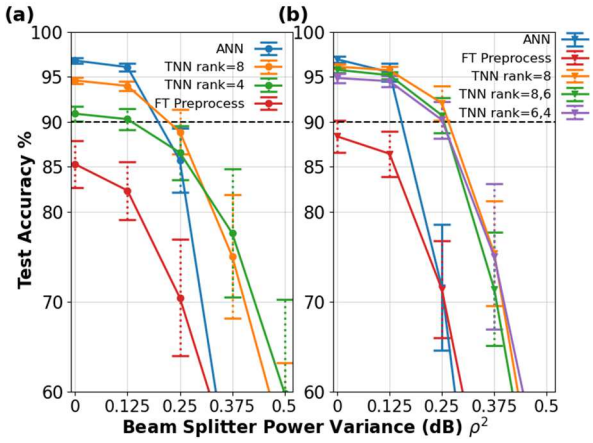

Fig. 7: Test Accuracy Percentage vs. Beam-Splitter power variation at $\boldsymbol{\sigma = 0.01}$ phase-shifter variance, **(a)** *case (a)* ANNs, and equivalent TNNs **(b)** *case (b)* ANNs and equivalent TNNs

Other hardware imprecisions can be the crosstalk between the phase-shifter controllers, the laser phase noise, unequal waveguide path lengths, etc. These imprecisions cannot be modeled as Gaussian noise due to their natures. For example, the laser phase noise will be more significant for MZIs located close to the output ports than the input ports. Thermal crosstalk and unequal path lengths will introduce deterministic impairments. In future works, these additional hardware imprecisions can be modeled in the simulations, an experiment can be conducted to demonstrate TNNs on a photonic integrated circuit.

Conclusion

In this work we simulated and demonstrated that a photonic TNN can still achieve *>90%* classification accuracy by using *33.6*× less MZIs than the conventional ANN, which can only achieve *71.6%*, under the practical hardware imprecisions such as phase-shifter variations and beam-splitter power imbalances.

This work was supported in part by AFOSR grant FA9550-181-1-0186.

Reference

[1] M. A. Nahmias, T. F. de Lima, A. N. Tait, H. Peng, B. J. Shastri, and P. R. Prucnal, "Photonic Multiply-Accumulate Operations for Neural Networks," *IEEE J. Sel. Top. Quantum Electron.*, vol. 26, no. 1, pp. 1–18, 2020.

[2] Y. Shen *et al.*, "Deep learning with coherent nanophotonic circuits," *Nat. Photonics*, vol. 11, no. 7, pp. 441–446, 2017.

[3] A. N. Tait *et al.*, "Neuromorphic photonic networks using silicon photonic weight banks," *Sci. Rep.*, vol. 7, no. 1, pp. 1–10, Dec. 2017.

[4] I. V. Oseledets, "Tensor-train decomposition," in *SIAM Journal on Scientific Computing*, Sep. 2011, vol. 33, no. 5, pp. 2295–2317.

[5] A. Novikov, D. Podoprikhin, A. Osokin, and D. P. Vetrov, "Tensorizing Neural Networks," in *Advances in Neural Information Processing Systems*, 2015, vol. 28.

[6] X. Xiao and S. J. Ben Yoo, "Tensor-Train Decomposed Synaptic Interconnections for Compact and Scalable Photonic Neural Networks," *2020 IEEE Photonics Conference (IPC)*, 2020, pp. 1-2.

[7] I. Williamson, T. W. Hughes, M. Minkov, B. Bartlett, S. Pai, and S. Fan, "Reprogrammable Electro-Optic Nonlinear Activation Functions for Optical Neural Networks," *IEEE J. Sel. Top. Quantum Electron.*, 2019.

[8] M. Reck, A. Zeilinger, H. J. Bernstein, and P. Bertani, "Experimental realization of any discrete unitary operator," *Phys. Rev. Lett.*, vol. 73, no. 1, pp. 58–61, Jul. 1994.

[9] W. R. Clements, P. C. Humphreys, B. J. Metcalf, W. S. Kolthammer, and I. A. Walmsley, "Optimal design for universal multiport interferometers," *Optica*, vol. 3, no. 12, pp. 1460–1465, 2016.

[10] A. Novikov, P. Izmailov, V. Khrulkov, M. Figurnov, and I. Oseledets, "Tensor Train Decomposition on TensorFlow (T3F)," 2020. Accessed: May 30, 2021. [Online]. Available: http://jmlr.org/papers/v21/18-008.html.

[11] J. R. Ong, C. C. Ooi, T. Y. L. Ang, S. T. Lim, and C. E. Png, "Photonic Convolutional Neural Networks Using Integrated Diffractive Optics," *IEEE J. Sel. Top. Quantum Electron.*, vol. 26, no. 5, Sep. 2020.

[12] M. Y.-S. Fang, S. Manipatruni, C. Wierzynski, A. Khosrowshahi, and M. R. DeWeese, "Design of optical neural networks with component imprecisions," *Opt. Express*, vol. 27, no. 10, p. 14009, May 2019.

[13] S. Pai, B. Bartlett, O. Solgaard, and D. A. B. Miller, "Matrix Optimization on Universal Unitary Photonic Devices," *Phys. Rev. Appl.*, vol. 11, no. 6, p. 1, 2019.