Digital-Analog Co-Design for Precision Compressed Integrated Photonic Convolution Neural Network

Yue Jiang, Wenjia Zhang* and Zuyuan He*

State Key Laboratory of Advanced Optical Communication Systems and Networks, Shanghai Jiao Tong University, Shanghai 200240, China. 200240

yue_jiang@sjtu.edu.cn, wenjia.zhang@sjtu.edu.cn, and zuyuanhe@sjtu.edu.cn

Abstract: Digital-Analog Co-design for photonic CNN with compressed precision is proposed to answer 3 practical concerns about analog precision: measurement and its mapping to digital domain, minimum demand of physical layer conditions and cheap methods to improve the performance of photonic CNN with poor precision.

1. Introduction

The ability to provide Tera-level computing power such as 11 T-FLOPS in [1] and 2 T-MACS reported in [2] has made photonic CNN architectures become a promising acceleration engine for artifical intelligence algorithms. However the harsh requirements of the 48 dB SNR to the physical layer in [1], challenging task of high-speed and multi-format modulation such as 50 Gbps PAM4 modluation and expensive ADC, DAC with 10-bit or higher resolution, which determine the precision of photonic CNN that can be deployed in practice is limtied to less than 8-bit.

Compared with computing platform such as TPU, GPU and FPGA, where the CNN algorithms running with Float 32, INT 16 or above, the operating precision is extremely compressed when loading pre-trained CNN in digital domain to the physical layer of photonic CNN in analog domain. Up to now, some efforts and analyses have been made to preserve the digital precision as much as possible, such as 8.5-bit weighting precision by designing the control circuit of micro-ring reported in [3] and 9-bit discussed in [4]. However, there is still a huge gap wait to bridge for photonic platform. Moreover, the damage to the performance induced by poor input precision and SNR conditions can't be ignored , which can hardly be solved with low price. Therefore, the solutions should be mining form both hardware level in analog domain and algorithms in digital domain namely Digital-Analog Co-Design for integrated photonic CNN, thus, some practical concerns are solved in this paper: The measurement of analog precision and its mapping relationship between datatype in digital domain, modeling of photonic CNN with limited precision, basic requirements to the physical layer to achieve acceptable performance and two potential algorithms to improve which under poor hardware conditions.

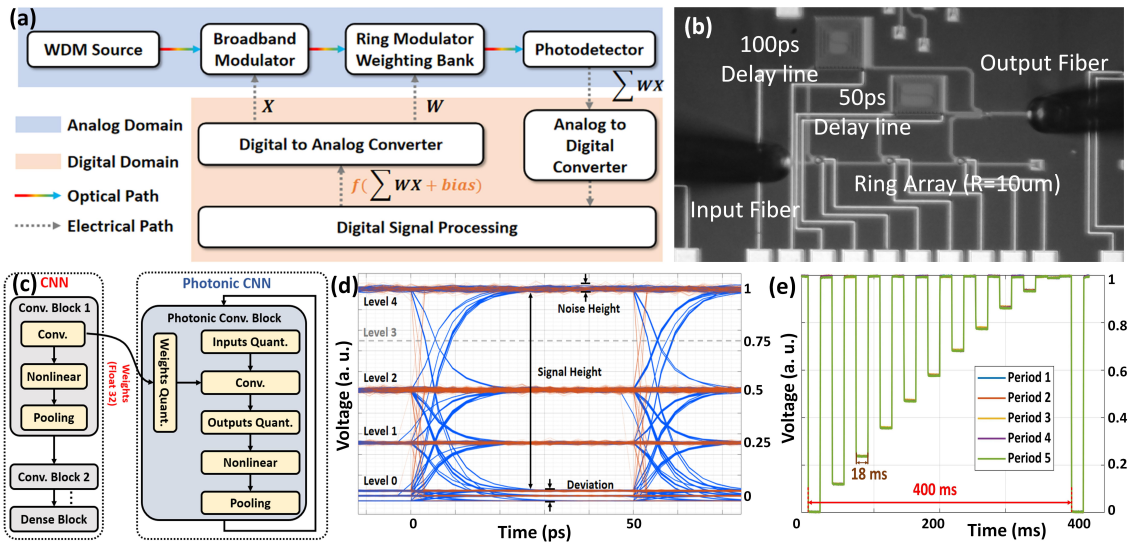

Fig. 1. (a) Setup of photonic CNN architecture based on frequency-time interleaved modulation. (b) Micro-Ring weighting bank. (c) Modeling of Photonic CNN with limited precision. (d) Eye diagram of 1-bit input vector X (0, 0.5, 1) mutiply with 1-bit weights W (0, 0.5, 1). (e) Precision measurement for micro-ring weighting bank.

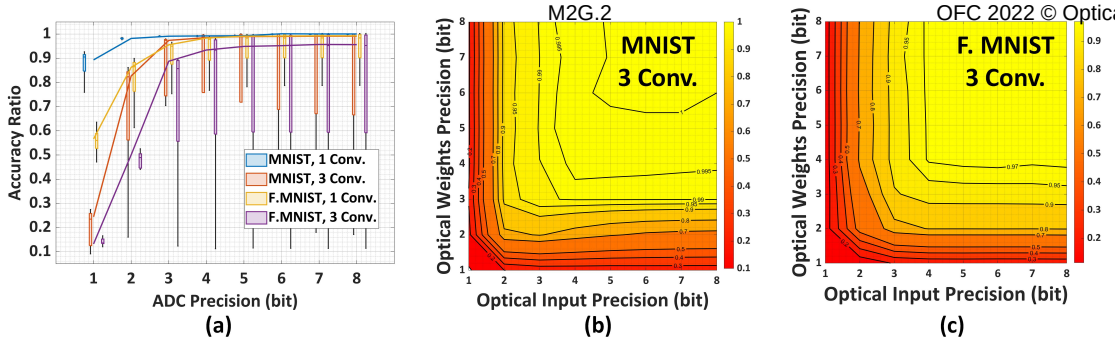

Fig. 2. (a) Performance evaluation for Photonic CNN in our work. (b) and (c) Test accuracy of photonic CNN with 8-bit ADC and 27 dB SNR under different P_w and P_x.

2. Modeling of Photonic CNN with Limited Precision

As shown in figure 1(a) and (b), we setup a photonic CNN architecture proposed in our previous work [5]. When the pre-trained weights W in digital domain is modulated to the intensity of optical signal in analog domain through the weighting banks driven by DAC, the original datatype of W such as INT 8 and Float 32 is mathematically quantified into UINT P_w, where P_w is the weighting precision in analog domain. The input vector X will also be quantified into UINT P_x when sent into photonic CNN through the broadband modulator (MZM in our work). As figure 1(d) shows, when X and W contains 3 levels (0, 0.5, 1) namely $P_x = 1$-bit, and $P_w = 1$-bit, the eye diagram of product WX will contain 5 linear distributed levels. The resolution of linear levels in eye diagram represents the real number range that can be accurately transmitted from analog domain to the digital domain, when levels=5, [-5, 5] in analog domain is equally to the datatype INT 3 or UINT 2. However, the level resolution is restrained by the SNR of optical circuit and precision of ADC, which indicates that quantization into INT $(P_c + 1)$ of convolution results is also implemented. Therefore, the physical layer equivalence model of photonic CNN is as shown in figure 1(c).

3. Digital-Analog Co-Design for Photonic CNN

3.1. Performance evaluation for physical layer conditions

The SNR of the whole optical circuit shown in figure 1 is 27 dB in our experiment, combine with ADC with 12-bit resolution, the maximum P_c is 8-bit. Meanwhile, the maximum P_x of MZM and P_w of weighting bank are measured by reading out the distinguishable linear levels in eye diagram or overlapped weighting curves shown in figure 1(e). Note that P_x, P_w will be always less than P_c, and P_x is decrease with the baud rate of the modulation, we finally achieved 8-bit P_x (3-bit at 10 Gbps) and 8-bit P_w.

Combine with the physical layer equivalence model and measurement of precision conditions, we can swiftly evaluate the performance of photonic CNN when loaded with various CNN model to deal with different tasks. Figure 2 showing the test accuracy ratio of photonic CNN loaded with one CNN and deep CNN for MNIST and Fashion MNIST under different ADC conditions, and the lines mark out the performance of our photonic CNN running at 10 Gbps in experiment. From another perspective, figure 2 provide the minimum requirements to the physical layer for two models and applications is P_w, P_x and P_c no less than 4-bit, corresponding with 13 dB SNR, which is much smaller than that calculated in [1] and more consistent with experimental results.

3.2. Performance improvement with customized algorithms

The mapping of analog precision to the datatype in digital domain make it possible to custom quantization algorithms in computer science [6, 7] for the photonic CNN to improve the performance. However, algorithms used in [7] cost expensive computing resources to solve the gradient disappearance and explosion induced by input and output quantization of photonic CNN. In this paper we proposed two cheap and potential methods.

3.2.1. Exhaust the robustness of the Dense block

The SNR requirements read out from figure 2 is much lower than that of theoretical calculation indicates that the dense block has high tolerance to make it possible compensating the penalty of the physical layer, as shown in figure 3(a), the pre-trained convolution block is deployed and fixed to the photonic CNN with equivalent precision of INT $(P_x + 1)$ and INT $(P_w + 1)$. The feature map extracted from the photonic CNN shown in figure 3(c) in analog domain can be processed as feature map in 64-bit computer added with the nosie induced by physical layer, whose pattern can be learned and compensated during the training process by feeding the photonic feature map into the dense block.

3.2.2. CNN based on edge extraction

As shown in figure 3(c), the feature map extracted by CNN have blurred when SNR=7 dB (2-bit), which leads to performance degradation of photon convolution neural network. Combine with the edge extraction technology,

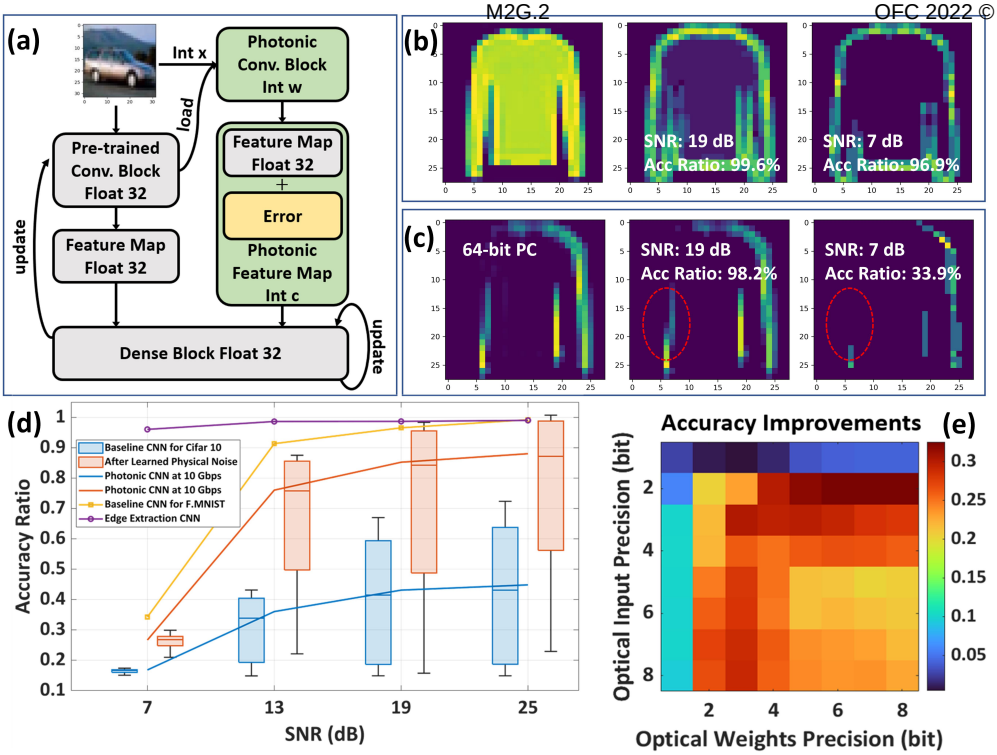

Fig. 3. (a) Physical noise learning in dense block. (b) Feature map extracted from T-shirt of Fashion MNIST by edge extraction CNN. (c) Feature map extracted from T-shirt of Fashion MNIST by baseline CNN. (d) performance compare of photonic CNN loaded with different models. (e) Accuracy improvements of after dense learned physical noise

the photonic CNN can learn the edge information of the object in images, which is still cognizable under poor precision conditions shown in figure 3(b).

These methods are quite simple and cheap for there are not extra operation layers or algorithms to assist during training and computing, as shown in figure 3(d), the performance of CNN learned noise pattern of physical are polished up under all SNR conditions and close to that of CNN in 64-bit computer when SNR $\geq$ 19 dB. figure 3(e) shows the advantages of the method in improving the performance of photonic CNN under the condition of $2 \leq P_w \leq$ 4-bit or $2 \leq P_x \leq$ 4-bit, which is suitable for practical photonic CNN. As for edge extraction CNN, it still maintain good performance of 96% accuracy ratio where as 34% of baseline CNN when loaded on the photonic device offering poor precision conditions less than 2-bit.

4. Conclusion

In this paper we discussed the Digital-Analog Co-Design for integrated photonic CNN. By reading out the SNR and resolution of linear levels in the eye diagram, the analog precision can be measured and mapped into datatype in digital domain, combine with the physical layer equivalence model, one can swiftly evaluate the performance and physical layer demands of photonic CNN, moreover, algorithms can be customized to improve the performance refer to the physical layer conditions, to achieve which, we also proposed two cheap but effective methods, by exhausting the robustness of dense block to learn physical layer noise, 30% gain of the test accuracy can be obtained, and CNN combined with edge extraction technique maintains 96.6% accuracy ratio under 2-bit conditions verse that of baseline CNN is 34.2%.

5. Acknowledgement

This work is supported by by National Natural Science Foundation of China under Grant 62175146 and The Major Key Project of PCL(PCL2021A14)

References

1. Xu, X. and Tan, M. and Corcoran, B. and Wu et al. Nature, 2021.
2. Feldmann, J. and Youngblood, N. and Karpov, M et al. Nature, 2021.
3. Zhang, W. and Huang, C. and Bilodeau, S et al. arxiV, 2021, https://arxiv.org/abs/2104.01164
4. Cheng, Qixiang and Kwon, Jihye and Glick, Madeleine and Bahadori et al. Proceedings of the IEEE, 2020.
5. Yue Jiang, Wenjia Zhang, Fan yang and Zuyuan He. Journal of Lightwave Technology, 2021.
6. J. Wu, C. Leng, Y. Wang, Q. Hu and J. Cheng. IEEE conference, CVPR, 2016.
7. Sunny, Febin P and Mirza, Asif and Nikdast, Mahdi and Pasricha, Sudeep, ACM TECS, 2021.