Photonic Max-Pooling for Deep Neural Networks Using a Programmable Photonic Platform

Farshid Ashtiani,[1,*,] Mehmet Berkay On,[1,2,*] David Sanchez-Jacome,[3] Daniel Perez-Lopez,[3] S. J. Ben Yoo,[2] and Andrea Blanco-Redondo[1]**

[1] *Nokia Bell Labs, 600 Mountain Ave, Murray Hill, NJ 07974, USA*
[2] *University of California Davis, Department of Electrical and Computer Engineering, One Shields Avenue, Davis, CA 95616, USA*
[3] *iPronics Programmable Photonics, Avenida Blasco Ibanez 25, Valencia 46010, Spain*
[*] *Equal contribution.*

[**] *farshid.ashtiani@nokia-bell-labs.com*

Abstract:
We propose a photonic max-pooling architecture for photonic neural networks which is compatible with integrated photonic platforms. As a proof of concept, we have experimentally demonstrated the max-pooling function on a programmable photonic platform consisting of a hexagonal mesh of Mach-Zehnder interferometers.

1. Introduction

Machine learning and deep neural networks have been essential to the recent advancements in a wide range of technologies, from communications and pattern recognition to medical diagnosis and treatment. As the number and variety of such applications increase, the need for more computation power grows. Digital clock-based systems, mainly graphics processing units (GPU), have been the dominating hardware platform for implementing artificial intelligence systems. Despite being highly reconfigurable and adaptable to different networks, their computation speed is generally limited by the clock frequency as well as the memory access time.

Photonic integrated circuits offer promising solutions to address the challenges with conventional digital platforms by benefiting from a very large optical bandwidth for signal processing as well as energy efficient interconnects. Different implementations of the essential blocks such as weight-and-sum for matrix multiplication (linear computation) and nonlinear activation functions have been demonstrated both optically and optoelectronically [1–6]. In addition to block-level, photonic neural networks have been demonstrated at the system level for applications such as image classification and fiber nonlinearity compensation [7, 8].

Deep neural networks typically consist of the interconnection of multiple layers of neurons with different configurations such as convolution, fully-connected, and pooling layers. Pooling layers are typically used to downsample a given matrix by calculating the average (average-pooling) or maximum (max-pooling) of its components (Fig. 1a). This results in invariance to spatial translations of the input, faster convergence, and less computation load for the next layers [9, 10]. Max-pooling layers are especially useful and widely used to extract bright features on dark backgrounds and can extract frequent as well as rare features [11]. Unlike average-pooling that is a linear transformation and can be implemented using linear components such as a multi-mode interferometer [12–14], max-pooling is a nonlinear transformation and implementing such function on a silicon photonic platform is challenging. In one demonstration, VCSELS is used as a nonlinear device [9] that, despite impressive results, is challenging to implement on integrated silicon photonic platforms. Here we propose an optical-input optical-output max-pooling architecture that benefits from the nonlinear characteristics of a ring modulator and is compatible with commercial silicon photonic processes. To experimentally demonstrate max-pooling functionality, the proposed architecture is implemented on the iPronics Smartlight Processor using a hexagonal mesh of Mach-Zehnder interferometers (MZI) [15].

2. Proposed Max-pooling Architecture

Figure 1b shows the proposed max-pooling architecture where A and B are the optical inputs and the output corresponds to the maximum of the inputs. Initially, the ring modulators are biased in such a way that their resonance wavelengths are symmetrically positioned around the wavelength of operation, λ_0 (Fig. 1c). The phase shifter within ring can be based on different tuning mechanisms such as carrier injection and thermal tuning, resulting in different processing speeds. The rings are followed by directional couplers to tap off a small fraction of the optical signals. The output of the couplers are coupled to two photodiodes (PD) connected in a balanced

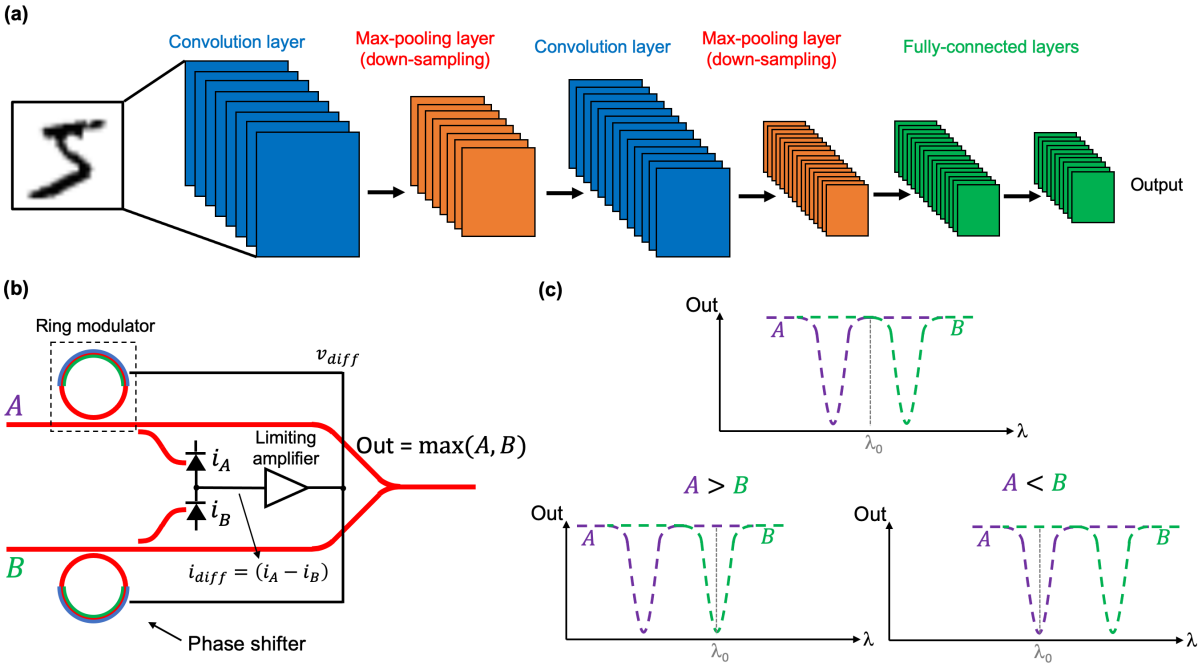

Fig. 1. (a) The structure of a typical deep neural network for image recognition and how max-pooling layers can be used to down-sample the input. (b) Proposed max-pooling architecture using integrated ring modulators and (c) the transfer characteristics for different input scenarios.

scheme. Therefore, the output current i_{diff} is a function of the difference in the input optical powers (*i.e.*, $A-B$). Then, i_{diff} is amplified using a limiting amplifier to drive the rings. The output voltage of the amplifier is limited to two values, *i.e.*, V_A and V_B, such that each voltage level brings the corresponding ring resonance to λ_0. If $A > B$ ($A < B$), the output current is positive (negative) and hence, v_{diff} equals V_A (V_B), aligining the resonance of ring B (A) with λ_0, heavily attenuating input B (A). In this case $|i_{diff}|$ is maximized, the system maintains this state for given inputs, and the output equals the larger optical input. Note that in either case, the resonance wavelength of one ring is locked to λ_0 making the system very stable and enables the detection of very small differences between the inputs which enhances the sensitivity and dynamic range of the circuit.

3. Experimental Results

As a proof of concept, the proposed max-pooling architecture is implemented on the iPronics Smartlight Processor. Note that this architecture can be implemented on any other integrated photonic platform with phase modulators and PDs. Figure 2a shows an MZI-based hexagonal mesh where the amplitude and phase of each arm can be controlled independently. The MZI mesh has one optical input that is split into two signals, A and B, using a tunable coupler to define the ratio between the two. On each arm, a ring modulator is formed by properly configuring the system where the optical phase within each ring can be controlled using a thermal phase shifter. The rings are identical and the resonance wavelengths are symmetrically biased around λ_0 = 1550 nm such that initially both rings have high transmissions. After each ring, 1% of the optical power is tapped off and photo-detected to generate the difference current that is amplified to drive the ring modulators. The two arms are then combined (red dashed box) and the output is monitored.

Figure 2b shows three different graphs. The green curve corresponds to the case that only the output of ring A is monitored while the input power ratio is swept using the input tunable coupler. The orange curve shows the similar scenario for ring B. In both cases, a monotonic increase in the output power can be seen. When the opto-electronic loop of Fig. 1b is closed and depending on the input power difference, the smaller input is heavily attenuated by its corresponding ring modulator when the resonance wavelength of the ring is aligned with 1550 nm. The blue dots in Fig. 2b show the measured output of the circuit as a function of the input power ratio (*i.e.*, $A/(A+B)$), which demonstrates the realization of max-pooling function (*i.e.*, output = $max(A,B)$). It can be seen that at any input power ratio, the output follows the larger input. Note that by using fast modulators (*e.g.,* PN-junction modulator) and wideband limiting amplifiers, bandwidths of tens of GHz (*i.e.*, picosecond response time) can be achieved.

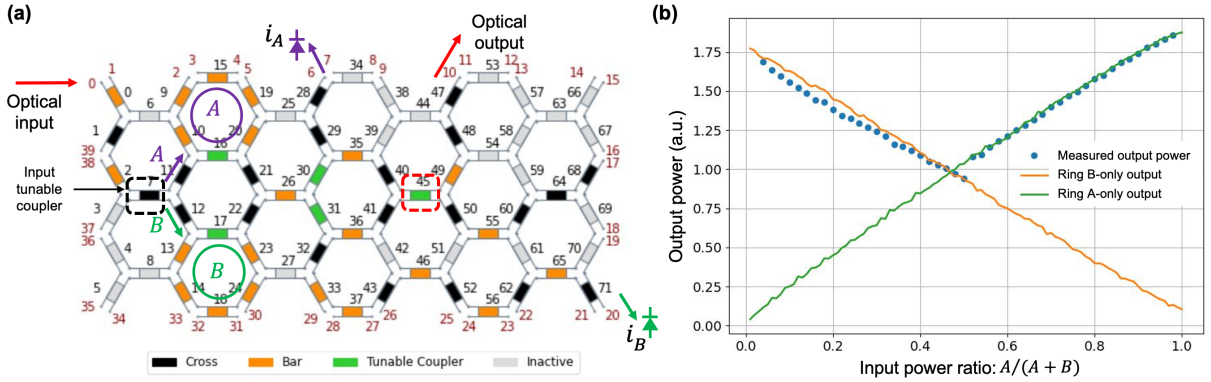

Fig. 2. (a) Implementation of the proposed max-pooling circuit on a programmable photonic platform. (b) Output power of the photonic max-pooling circuit as a function of the input power ratio.

4. Summary and Conclusion

A photonic max-pooling architecture is proposed where two input optical signals are coupled to two ring modulators. The resonance wavelength of the ring corresponding to the smaller input is locked to the wavelength of operation resulting in large attenuation of the smaller signal while the larger input is transmitted to the output with no attenuation. The functionality of the circuit is experimentally demonstrated using a programmable photonic platform. For larger scale networks, the number of optical inputs can be increased by using multiple 2-input blocks in parallel. Moreover, the proposed architecture can be implemented on other integrated photonic platforms and by using fast modulators and PDs, an operation bandwidth of tens of GHz can be achieved.

References

1. J. Feldmann, et al. "Parallel convolutional processing using an integrated photonic tensor core," *Nature* **589**, 52–58 (2021).
2. H. Zhou, et al. "Photonic matrix multiplication lights up photonic accelerator and beyond," *Light Sci Appl* **11**, 30 (2022).
3. A.N. Tait, et al. "Neuromorphic photonic networks using silicon photonic weight banks," *Sci Rep* **7**, 7430 (2017).
4. A. Jha, C. Huang, and P. R. Prucnal, "Reconfigurable all-optical nonlinear activation functions for neuromorphic photonics," *Opt. Lett.* **45**, 4819-4822 (2020).
5. R. Amin, et al., "ITO-based electro-absorption modulator for photonic neural activation function," *APL Materials* **7**, 081112 (2019).
6. A. N. Tait, T. F. de Lima, M. A. Nahmias, H. B. Miller, H. Peng, B. J. Shastri, and P. R. Prucnal, "Silicon photonic modulator neuron," *Phys. Rev. Applied* **11**, 064043 (2019).
7. F. Ashtiani, A.J. Geers, and F. Aflatouni, "An on-chip photonic deep neural network for image classification," *Nature* **606**, 501–506 (2022).
8. C. Huang, et al. "A silicon photonic–electronic neural network for fibre nonlinearity compensation," *Nat Electron* **4**, 837–844 (2021).
9. Y. Zhang, S. Xiang, X. Guo, A. Wen and Y. Hao, "The Winner-Take-All Mechanism for All-Optical Systems of Pattern Recognition and Max-Pooling Operation," in *Journal of Lightwave Technology* **38**, 5071-5077 (2020).
10. D. Dang, S. V. R. Chittamuru, S. Pasricha, R. Mahapatra, and D. Sahoo, "BPLight-CNN: A photonics-based backpropagation accelerator for deep learning," *J. Emerg. Technol. Comput. Syst.* **17**, 4, Article 49 (October 2021)
11. N. Murray and F. Perronnin, "Generalized Max Pooling" *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 2473-2480.
12. L. De Marinis, F. Nesti, M. Cococcioni, and N. Andriolli, "A Photonic Accelerator for Feature Map Generation in Convolutional Neural Networks," in *OSA Advanced Photonics Congress (AP)* OSA Technical Digest (Optica Publishing Group, 2020), paper PsTh1F.3.
13. V. Bangari et al., "Digital Electronics and Analog Photonics for Convolutional Neural Networks (DEAP-CNNs)," in *IEEE Journal of Selected Topics in Quantum Electronics* **26**, 1-13 (2020).
14. E. Paolini, et al. "Photonic-aware neural networks," *Neural Comput & Applic* **34**, 15589–15601 (2022)
15. D. Pérez, et al. "Multipurpose silicon photonics signal processor core," *Nat Commun* **8**, 636 (2017).