Activation Stretching for Tackling Noise in Photonic Aware Neural Networks

E. Paolini,[1,2,3] L. De Marinis,[1] L. Valcarenghi,[1] L. Maggiani,[3] and N. Andriolli[2]

[1] *Scuola Superiore Sant'Anna, Pisa, 56124, Italy (emilio.paolini@santannapisa.it)*
[2] *Consiglio Nazionale delle Ricerche (CNR-IEIIT), Pisa, 56122, Italy*
[3] *Sma-RTy Italia SRL, Carugate, 20061, Italy*

Abstract: This paper introduces a stretching strategy for nonlinear activation functions aimed to enhance noise resilience in photonic-aware neural networks. Its effectiveness is numerically demonstrated in counteracting different noise levels in low-resolution operations.

1. Introduction

Over the past few years, Neural Networks (NNs) models have achieved impressive results in numerous practical applications, yet these outstanding achievements required a substantial rise in model complexity, leading to a large amount of floating-point operations (FLOPs) in both training and inference processes [1]. Since traditional hardware cannot keep up with this trend [2], research efforts have focused on both software solutions, e.g., NN model shrinking, and hardware solutions, e.g., computing platforms that can emulate NN structures, improving speed and power consumption [1].

Among hardware-based solutions, photonic neural networks are at the forefront of scientific and industrial research, leveraging the high bandwidth, low latency, and low power consumption of photonics [3]. Several photonics-based solutions for accelerating NN computations have been proposed. In [4], a photonic Multiply-Accumulate (MAC), namely PMAC, is proposed relying on reconfigurable photonic elements, i.e., Mach–Zehnder interferometer, micro-ring resonators, and directional couplers. In [5] a noise-resilient coherent photonic NN is experimentally demonstrated, performing classification on the MNIST dataset at 20 GMAC/s. A hybrid photonic-electronic neuron is proposed in [6]: this architecture is designed to leverage the strengths of both optical and electrical approaches. Co-simulations show that it can provide a hundredfold increase in the number of connections per neuron, while working at tens of GMAC/s.

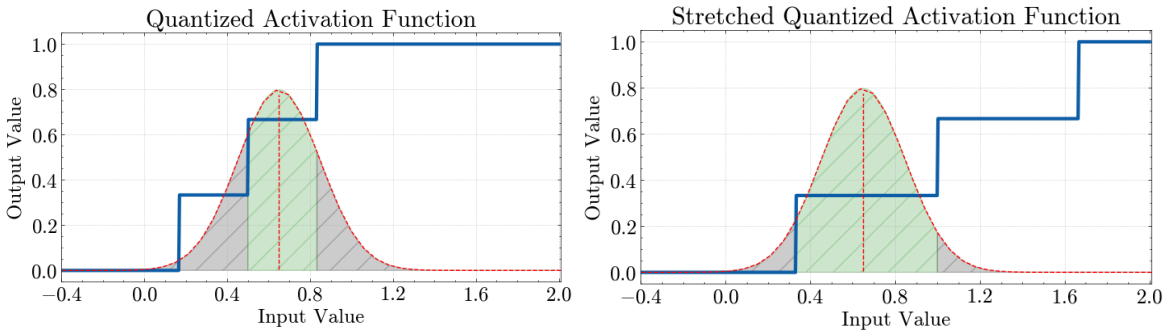
Although photonic neuromorphic devices promise to replace electronic processors for NN inference acceleration, their physical layer issues hinder a widespread adoption in real world problems. The limitations deriving from the photonic hardware can be categorized in architectural and resolution constraints. Concerning the first group, the limitations can be summarized as: (i) positive-valued inputs, as they are encoded in the intensity of the optical signals; (ii) maximum number of inputs, namely fan-in, to each neuron, ≈ 10 in all-optical approaches and ≈ 200 in opto-electronic implementations. The second type of constraint, i.e., resolution, arises from the presence of noise and distortions in the analog optical hardware, limiting the number of distinguishable values [7]. Hence, typical resolutions for photonic architectures range from 1 to 6 bits after analog-to-digital conversion [8]. These constraints require a proper NN model design and training, motivating the introduction of a class of ad-hoc NN models compliant with underlying physical constraints, referred to as Photonic-Aware Neural Network (PANN) [7]. Additionally, noise sources affect different elements of photonic NNs, i.e., inputs, weights, and activation functions. As proposed in [9], counteracting noise in the training process of NNs improves their performance, obtaining models robust to various physical distortions arising from underlying photonic hardware.

In this work, we further extend the PANN concept, introducing a stretching strategy for activation functions aimed to improve the noise resilience in low-resolution photonic NNs. Different noise levels, due to different photonic hardware, can be tackled by stretching the activation function accordingly, resulting in more effective PANNs.

2. Activation Stretching

Counteracting the noise introduced by analog hardware is paramount for the development of high-accuracy photonic NNs. To this aim, we propose a stretching strategy that minimizes the impact of different noise levels arising from the photonic hardware.

When considering noise, the output of a layer is represented by a probability distribution. As discussed in [9], the noise sources of the basic building blocks of a photonic neuron can be modeled as Additive White Gaussian



(a) (b)

Fig. 1: Quantization interval without activation stretching (a) and when a stretching $\alpha = 2$ is applied (b).

Noise (AWGN). Hence, the output of a linear part of the layer can be defined as: $\Pi = f(w,x) + \mathcal{N}(0,\sigma^2)$, with f representing the layer linear operations, w and x the arbitrary precision weights and inputs, respectively, and $\mathcal{N}(0,\sigma^2)$ the AWGN process with zero average and σ^2 variance. Focusing on the training phase, [10] showed that the direct quantization of floating-point parameters results in high accuracy loss, hence a quantization-aware training scheme should be exploited, as proposed in [7]. Thus, the output of the layer is defined as:

$$y = \varphi(f(q_{\text{kernel}}(w), q_{\text{input}}(x)) + \mathcal{N}(0,\sigma^2))$$

with q_{kernel}, q_{input} representing the weight and input quantizers, and φ the quantized activation function. The result of the application of φ when considering the AWGN is depicted in Fig. 1a for a 2-bit quantized ReLU. The layer output may remain in the proper quantization interval (green area in the figure), or it may fluctuate to an undesired quantized interval (grey area), resulting in a wrong computation. To overcome this issue, we propose a stretching strategy in which the quantized activation function is adjusted taking into account the noise due to the photonic hardware. Specifically, the stretching factor α is applied to the activation function, with the output of the layer defined as:

$$y = \varphi\left(\frac{f(q_{\text{kernel}}(w), q_{\text{input}}(x)) + \mathcal{N}(0,\sigma^2)}{\alpha}\right)$$

Thus, by stretching the quantization function, as shown in Fig. 1b for $\alpha = 2$, the probability to fall into the correct quantization interval increases, even in the presence of noise. Indeed, the green area in Fig. 1b is wider compared to the one sketched in Fig. 1a. Furthermore, it is worth noting that adjusting the quantization intervals of the activation function does not impact the final accuracy because the quantization-aware scheme, applied during training to mitigate quantization-induced losses, effectively accounts for this modification [7].

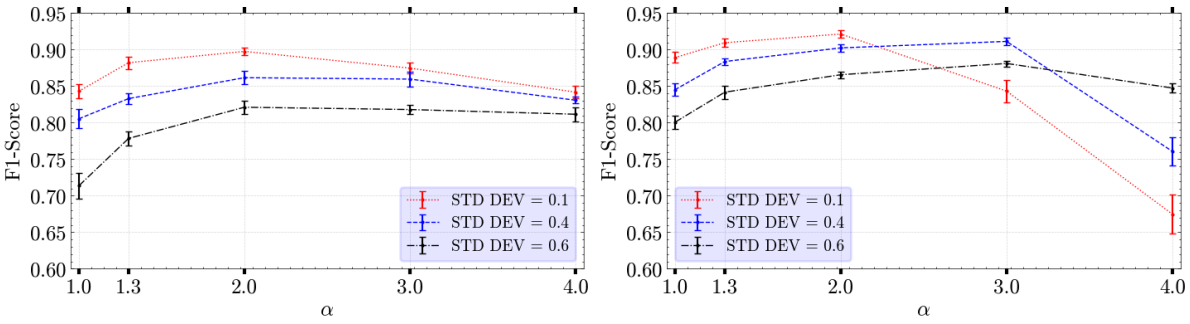
3. Experiments and Results

To demonstrate the effectiveness of the proposed stretching strategy, we trained different PANNs considering three noise levels, i.e., standard deviation of the AWGN, $\sigma = \{0.1, 0.4, 0.6\}$, and two bit resolutions of the analog-to-digital and digital-to-analog conversion hardware, i.e., 2 and 4 bits. The experiments have been repeated for different values of α, i.e., 1, 1.3, 2, 3, 4. The PANN model has been developed for image recognition on the Street View House Numbers (SVHN) dataset [11], and is composed of 8 convolutional layers with 3×3 of kernel size, and 2 fully-connected layers, composed of 64 and 10 neurons, respectively. The SVHN dataset contains 10 classes and input images of 32×32 pixels on 3 channels, i.e., RGB. Since the dataset is not perfectly balanced, a class weighting method has been applied during training and the F1-score is used to evaluate the performance of the model, with confidence intervals at 95% confidence level computed by repeating 10 times the training procedure.

F1-scores as a function of α are reported in Fig. 2a and Fig. 2b for 2-bit and 4-bit resolutions, respectively.

In the 2-bit scenario, for all considered noise levels, the highest F1-score is reached when $\alpha = 2$: 0.90, 0.86, and 0.82 for $\sigma = 0.1$, 0.4, and 0.6, respectively. Hence, by setting $\alpha = 2$, the F1-score increase with respect to the no-stretching scenario, i.e., $\alpha = 1$, reaches 5.5%, 5.6%, and 10.7% for $\sigma = 0.1$, 0.4, and 0.6, respectively. In particular, the performance gain increases with the noise magnitude. Additionally, when α increases too much the F1-score starts decreasing, showing that an optimal value of α exists.

When considering 4-bit experiments, reported in Fig. 2b, slightly higher F1-scores are obtained due to the larger resolution, i.e., 0.92, 0.91, and 0.88 in the best case for $\sigma = 0.1$, 0.4, and 0.6, respectively. Also in this case, an optimal α exists, however it depends on the considered noise magnitude: for $\sigma = \{0.1\}$, the optimal α value is 2, while for $\sigma = 0.4, 0.6$, the best F1-score is obtained with $\alpha = 3$. The F1-score increase with respect to the



(a) 2-bit. (b) 4-bit.

Fig. 2: F1-score as a function of α for different bit resolutions.

no-stretching scenario is 3.2%, 6.6%, and 8.1% for $\sigma = 0.1$, 0.4, and 0.6, respectively, highlighting the stretching effectiveness especially at higher noise levels, as in the 2-bit scenario.

4. Conclusions

Photonic NNs promise to outperform electronic counterparts in terms of speed and power consumption. However, several limitations still hamper their application in real use-cases. Among these, the presence of noise in optical devices represents one of the main issues. Hence, with the aim of limiting the impact of noise in NN computations, in this work we have proposed an activation stretching strategy to improve PANN performance. Specifically, the proposed method proved to be effective in counteracting noise, especially in high noise magnitude. Indeed, the maximum F1-score increase occurs for $\sigma = 0.6$ (10.7% in the 2-bit case and 8.1% in the 4-bit case). In conclusion, the stretching strategy effectively addresses noise in PANNs, paving the way for effective photonic neuromorphic computations.

5. Acknowledgments

This work is partly supported by the project CLEVER (project number 101097560). The project is supported by the Key Digital Technologies Joint Undertaking and its members (including top-up funding by the Italian MUR), and by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on “Telecommunications of the Future” (PE00000001 - program “RESTART”).

References

1. K. Liao, T. Dai, Q. Yan, X. Hu, and Q. Gong, “Integrated Photonic Neural Networks: Opportunities and Challenges,” ACS Photonics (2023).
2. A. Cem, S. Yan, Y. Ding, D. Zibar, and F. Da Ros, “Data-driven modeling of mach-zehnder interferometer-based optical matrix multipliers,” J. Light. Technol. (2023).
3. G. Giamougiannis, A. Tsakyridis, M. Moralis-Pegios, C. Pappas, M. Kirtas, N. Passalis, D. Lazovsky, A. Tefas, and N. Pleros, “Analog nanophotonic computing going practical: silicon photonic deep learning engines for tiled optical matrix multiplication with dynamic precision,” Nanophotonics **12**, 963–973 (2023).
4. A. Mosses and P. J. Prathap, “Design and analysis of on-chip reconfigurable photonic components for photonic multiply and accumulate operation,” Opt. Quantum Electron. **55**, 934 (2023).
5. G. Mourgias-Alexandris, M. Moralis-Pegios, A. Tsakyridis, S. Simos, G. Dabos, A. Totovic *et al.*, “Noise-resilient and high-speed deep learning with coherent silicon photonics,” Nat. Commun. **13**, 5572 (2022).
6. L. De Marinis, A. Catania, P. Castoldi, G. Contestabile, P. Bruschi, M. Piotto, and N. Andriolli, “A codesigned integrated photonic electronic neuron,” IEEE J. Quantum Electron. **58**, 1–10 (2022).
7. E. Paolini, L. De Marinis, M. Cococcioni, L. Valcarenghi, L. Maggiani, and N. Andriolli, “Photonic-aware neural networks,” Neural Comput. Appl. **34**, 15589–15601 (2022).
8. B. Shi, N. Calabretta, and R. Stabile, “Deep neural network through an InP SOA-based photonic integrated cross-connect,” IEEE J. Sel. Top. Quantum Electron. **26**, 1–11 (2019).
9. M. Kirtas, N. Passalis, G. Mourgias-Alexandris, G. Dabos, N. Pleros, and A. Tefas, “Robust architecture-agnostic and noise resilient training of photonic deep learning models,” IEEE TETCI **7**, 140–149 (2022).
10. E. Paolini, L. De Marinis, M. Cococcioni, L. Valcarenghi, L. Maggiani, and N. Andriolli, “Photonic-Aware Neural Network: a fixed-point emulation of photonic hardware,” in *Proc. OECC/PSC 2022,* .
11. Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, “Reading Digits in Natural Images with Unsupervised Feature Learning,” in *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011,* (2011).