Frontiers in Optics + Laser Science 2024 (FiO/LS) © Optica Publishing Group 2024

Variational Inference Using High Speed Photonic Neural Networks

James Garofolo,[1] Taichu Shi,[1] Paul Prucnal,[2] and Ben Wu[1,*]

[1]*Department of Electrical and Computer Engineering, Rowan University, Glassboro, NJ 08028, USA*
[2]*Lightwave Communications Laboratory, Department of Electrical Engineering, Princeton, NJ 08544*
wub@rowan.edu

Abstract: We propose a method of performing variational inference using high speed photonic neural accelerators. This method incurs no slowdown compared to deterministic photonic inference, affecting only the power consumption of existing accelerator architectures.

OCIS codes: 200.4260 Neural networks; 200.4700 Optical neural systems

1. Introduction

Bayesian Neural Networks (BNNs) are a powerful type of model in deep learning due to their ability to estimate epistemic uncertainty. Through the training process described in [1], they learn to minimize predictive variance on in-distribution samples, leaving anomalous samples measurably more uncertain. Unfortunately, this technique is limited in usability due to the time needed to approximate the mean and variance integrals of a Gaussian distribution. Novel methods of performing this approximation, called Variational Density Propagation (VDP) methods, can reduce time complexity [2], but BNN forward passes remain slower than deterministic inference.

Photonic neural network accelerators are an emerging technology that increases the throughput and decreases the latency and power consumption of neural inference over digital computers [3]. These advantages come at the cost of a lower computational precision [4], which is seen as an acceptable trade-off when working with approximation models. That said, with weight uncertainty being desirable for BNNs, photonic neural accelerators show promise for implementing them in continuous time. Previous works have explored this [5], but was only able to propagate uncertainty through matrix multiplications as the computations were done on separate circuits.

Here we propose a method of performing VDP in the continuous time domain with circuits made to perform deterministic photonic neural inference. This method leverages unused bandwidth above the 3dB cutoff of the sampling circuitry of the system to convey noise at frequencies that would accumulate statistical significance over the course of a single inference step. The central tendency and variability are measured using frequency domain operations, with high frequency noise adding uncertainty to low frequency mean signals. This method allows for propagation through analog activation functions before sampling, and considers inherent circuit noises in the measurement of signal variability. The system used to perform the proposed operation is shown in Fig. 1a.

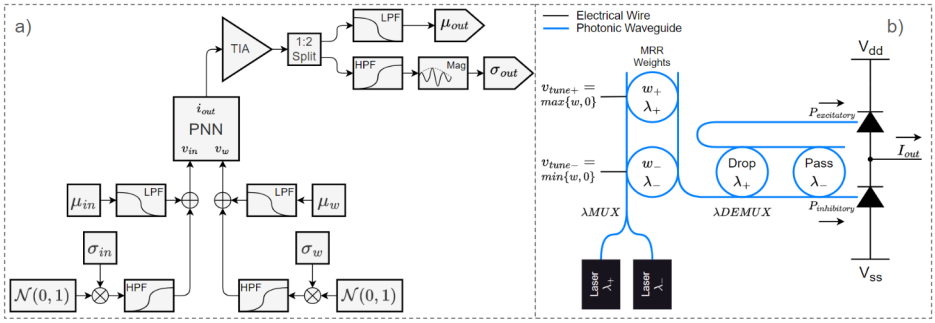

Fig. 1. a) Block diagram of the proposed photonic BNN b) Schematic of the simulated photonic circuit, adapted with permission from [6]

2. System Setup

The simulated circuit used to validate this method is shown in Fig. 1b. It uses two PIN junction ring modulators feeding constant weight through-drop resonators to form a two-quadrant multiplier. The electrical transfer functions for these modulators were fitted to match those measured in [6]. Excitatory and inhibitory inputs are fed through separate wavelengths, modulated off-chip and constrained to be positive by the use of ReLU activations in the implemented neural network. The signal means were transmitted at a rate of 5GBd, consistent with [3]. Additive Gaussian noise was conveyed using the frequency range from 7GHz to 20GHz, with the signal power below 7GHz removed to ensure a cleanly transmitted signal mean, save for inherent circuit noises. The resulting output noise was removed using a low-pass filter to recover the mean, and measured using a high-pass filter and magnitude detector to approximate the variance, as illustrated in Fig. 1a.

To verify the proposed method, the proposed stochastic multiplication method was simulated with direct digital synthesis (DDS) implementing a fully connected neural network using the PyTorch machine learning library [7]. This model was trained, along with a control model using digital extended VDP [2], to classify samples from the MNIST handwritten digits dataset [8]. Both models were trained using the same hyperparameters for 25 epochs,

and the weight tensors that performed best on the validation dataset were saved for evaluation. The models' anomaly detection behavior was then examined by perturbing the test dataset with additive white Gaussian noise at various signal-to-noise ratios and computing the average predictive variance for each severity.

3. Results and Analysis

Photonic VDP was trained to optimality after 14 epochs, reaching a validation accuracy of 95.70% on unperturbed handwritten digits. ExVDP was trained to optimality after 15 epochs, reaching a validation accuracy of 96.41%. The results of the Gaussian perturbation experiment are shown in Fig. 2. As these figures show, the models performed near identically, with slightly less severe change as signal-to-noise ratio increases for Photonic VDP. This behavior can be improved by increasing the electrical bandwidth of the noise, and thus the statistical significance of the approximation. This experiment was limited by the sample rate of the simulated DDS system.

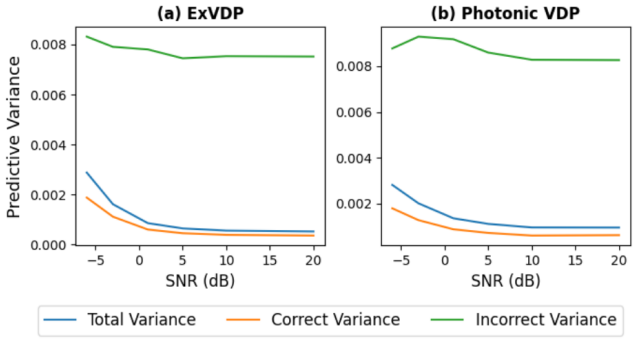

Fig. 2. Average predictive variance for all predicted samples (blue), correctly classified samples (orange) and misclassified samples (green) versus signal-to-noise ratio for a) Extended VDP, as proposed by [2]. b) The proposed photonic VDP method.

This method is theoretically just as fast as deterministic photonic neural acceleration, which has been shown to outperform early GPU architectures on deterministic convolution by speedup ratios from 1.4x to 7x [3]. The added benefit of uncertainty estimation comes at the cost of power consumption, as the proposed method effectively doubles the required number of digital-to-analog and analog-to-digital converters. Using the estimation criteria cited in [3], these additions should increase the power consumption of the circuit by around 1.6x, which is still below the power consumption of the GPU architectures used to benchmark the original design.

4. Conclusion

We propose a method of implementing BNNs with high speed photonic neural accelerators. This method is just as fast as deterministic photonic inference, increasing the power consumption by roughly 1.6x while still remaining more energy-efficient than digital inference. The anomaly detection behavior is shown to be near equally as effective as modern digital BNNs, and can be improved further by way of using programmable analog noises over DDS. This work was supported by the National Science Foundation (NSF) under Grant ECCS- 2128608.

References

1. Giuseppina Carannante, Dimah Dera, Orune Aminul, Nidhal C. Bouaynaya, and Ghulam Rasool. Self-assessment and robust anomaly detection with bayesian deep learning. In *2022 25th International Conference on Information Fusion (FUSION)*, pages 1–8, 2022.
2. Dimah Dera, Ghulam Rasool, and Nidhal Bouaynaya. Extended variational inference for propagating uncertainty in convolutional neural networks. In *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2019.
3. Viraj Bangari, Bicky A. Marquez, Heidi Miller, Alexander N. Tait, Mitchell A. Nahmias, Thomas Ferreira de Lima, Hsuan-Tung Peng, Paul R. Prucnal, and Bhavin J. Shastri. Digital electronics and analog photonics for convolutional neural networks (deap-cnns). *IEEE Journal of Selected Topics in Quantum Electronics*, 26(1):1–13, 2020.
4. Weipeng Zhang, Chaoran Huang, Hsuan-Tung Peng, Simon Bilodeau, Aashu Jha, Eric Blow, Thomas Ferreira de Lima, Bhavin J. Shastri, and Paul Prucnal. Silicon microring synapses enable photonic deep learning beyond 9-bit precision. *Optica*, 9(5):579–584, May 2022.
5. Changming Wu, Xiaoxuan Yang, Yiran Chen, and Mo Li. Photonic bayesian neural network using programmed optical noises. *IEEE Journal of Selected Topics in Quantum Electronics*, 29(2: Optical Computing):1–6, 2023.
6. Weipeng Zhang, Joshua C. Lederman, Thomas Ferreira de Lima, Jiawei Zhang, Simon Bilodeau, Leila Hudson, Alexander Tait, Bhavin J. Shastri, and Paul R. Prucnal. A system-on-chip microwave photonic processor solves dynamic rf interference in real time with picosecond latency. *Light: Science amp; Applications*, 13(1), Jan 2024.
7. Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703, 2019.
8. Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.