# Error Resilient Neuromorphic Systems Using Embedded Predictive Neuron Checks

Chandramouli Amarnath and Abhijit Chatterjee
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332–0250
Email: chandamarnath@gatech.edu, abhijit.chatterjee@ece.gatech.edu

Abstract—The reliability of emerging neuromorphic compute fabrics is of great concern due to their widespread use in critical data-intensive applications. Ensuring such reliability is difficult due to the intensity of underlying computations (billions of parameters), errors induced by low power operation and the complex relationship between errors in computations and their effect on network performance accuracy. We study the problem of designing error-resilient neuromorphic systems where errors can stem from: (a) soft errors in computation of matrix-vector multiplications and neuron activations, (b) malicious trojan and adversarial security attacks and (c) effects of manufacturing process variations on analog crossbar arrays that can affect DNN accuracy. The core principle of error detection relies on embedded predictive neuron checks using invariants derived from the statistics of nominal neuron activation patterns of hidden layers of a neural network. Algorithmic encodings of hidden neuron function are also used to derive invariants for checking. A key contribution is designing checks that are robust to the inherent nonlinearity of neuron computations with minimal impact on error detection coverage. Once errors are detected, they are corrected using probabilistic methods due to the difficulties involved in exact error diagnosis in such complex systems. The technique is scalable across soft errors as well as a range of security attacks. The effects of manufacturing process variations are handled through the use of compact tests from which DNN performance can be assessed using learning techniques. Experimental results on a variety of neuromorphic test systems: DNNs, spiking networks and hyperdimensional computing are presented.

Index Terms—Neural Networks, Fault Tolerance, Resilience

## I. Introduction

The increasing complexity and capability of neuromorphic systems has resulted in increased growth of required compute capacity for running these systems. Fig. 1 details the two major phases of FLOPs deployed on state-of-the-art AI systems - initially doubling every two years, but later doubling every 2-4 months (doubling every 2 months for language processing applications) [1]. This increasing FLOP requirement translates to larger energy usage (and therefore carbon emissions) for training and running these neuromorphic systems. Patterson et al [2] have reported that running training and neural architecture search for a large, modern transformer model emits more than 600,000 tons of $CO_2$ (compared to an average of 126,000 tons emitted over the lifetime of a single car). However, the use of sparsity in activation, low supply voltages,

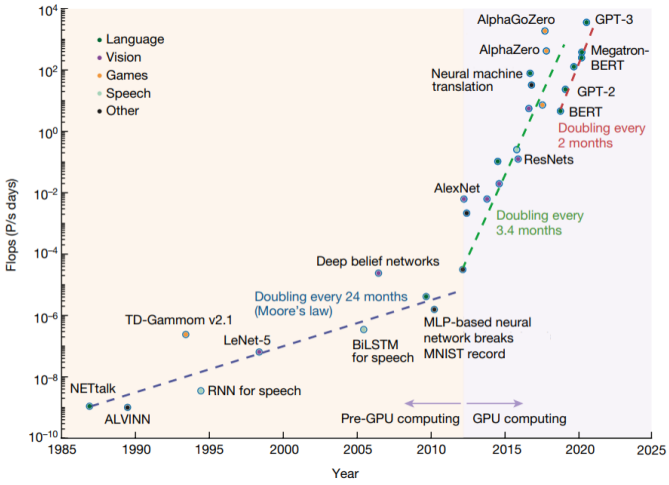


Fig. 1: FLOPs demanded by AI applications per year ( [1])

novel devices and 3D integration technologies can help an mitigate this excessive energy consumption [1], [3]. However, the use of low supply voltages along with coupled noise due to dense device integration, and use of novel devices, accelerates bit errors in DNN computations and accelerator memory. Of key concern is the impact of such errors on DNN inference accuracy [4]. One approach is fault and error-aware training of DNNs. This is very simulation-intensive especially for large DNNs. Moreover, analog RRAM based DNNs suffer from the effects of manufacturing process variations. These are particularly vulnerable to systematic process variability effects [5].

In the following, we present an overview of an error resilience framework for neuromorphic systems, relying on dimensionality reduction (encoding) methods for on-line error resilience. We also leverage computation of neuron output gradients for error detection and suppression. For testing of analog RRAM based DNNs that are impacted by manufacturing process variations, we propose an alternative test scheme that does not require application of a complete validation test dataset to the DNN for determining its accuracy, thus reducing test costs dramatically.

## II. Overview

Error resilience techniques discussed below for neuromorphic systems have two components: (1) detection and correc-

979-8-3503-2597-3/23/31.002023IEEE

2023 IEEE 24th Latin American Test Symposium (LATS) | 979-8-3503-2597-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/LATS58125.2023.10154498

tion of soft errors in DNN computations and memory accesses and (2) testing and tuning of DNNs under random as well as systematic manufacturing process variability effects.

Online Error Resilience: Error detection in DNN MAC computations has been accomplished with low overhead and high coverage using encoded checksums [6]. Our approach extends error detection to activation computations using regressors to predict compressed (encoded) layer outputs from compressed (encoded) layer inputs, thresholding for error detection based on the nominal distribution of the prediction error across training data. This achieved high error coverage and low false alarms for activation errors with low overhead ( 2-4% FLOP overhead) [7].

Error resilience in DNN computations has been examined in prior work, using median feature selection to filter out extreme erroneous values (altering training to allow median feature selection to work) [8]. Ranger [9] clamped the outputs of DNN activation functions to their maxima and minima over training data to prevent errors from causing misclassifications, achieving high effectiveness at low overhead. However, these approaches typically suffer from performance degradation for weight parameter errors and may require modification of DNN training ( [8]).

We also leverage neuron output distributions across training data to establish correlations and thresholds for error suppression, allowing fast, low-overhead online error resilience in neuromorphic systems. Our approach leverages the differences (gradients) in neuron outputs between neighboring neurons across training data to establish statistical thresholds for error localization, followed by error suppression (zeroing erroneous neuron values), achieving superior performance to prior work with comparable overhead [10]. Similar approaches were used for transformers, checking for out-of-range values (based on training data neuron values) at each computation in the transformer and setting those values to zero (suppression)s, showing near-total effectiveness for weight parameter errors and soft errors in transformers. Similar suppression systems were also tested for spiking neural networks [5], suppressing input spike trains that flagged as erroneous based on thresholds established across training data, allowing up to 60% accuracy improvement compared to the baseline no-resilience case. The application of statistical thresholding using on-line statistics calculated during training allows the use of these approaches in reinforcement learning, achieving near-total effectiveness on the Atari Pong DQN benchmark.

Offline Testing: Testing RRAM analog crossbar neuromorphic devices using the entire test dataset for each device can be expensive for large test datasets. Our approach aims to compress the test dataset to a representative test ensemble, capturing correlations between the DNN response to the compressed test set (test response ensemble) and the classification accuracy using regressors. In this way the classification accuracy of the DNN can be predicted from a compact test set (subset of the test dataset). This captures and quantifies variations and nonidealities in the DACs (integral nonideality), RRAM devices (conductance variations due to random and systematic manufacturing process variations and operating temperature) and quantization errors in ADCs. RRAM variability is modeled, via Spice RRAM device simulation, by a multiplicative weight perturbation coefficient, that modulates the coefficients of RRAM matrix dot product computations to reflect RRAM array nonidealities. Prediction of performance from compressed subsets of the test dataset using a trained regressor, allows test speedups of up to 20x on VGG16 trained on CIFAR10 [11].

Future Work includes exploration of tuning and reconfiguration of neuromorphic hardware using the results of fast offline testing to improve device yield. These approaches can further be extended to multimodal AI, hyperdimensional computing, continual online learning and to advanced technologies such as FeRAMs, MRAMs and 3-D integration, due to the generalizable nature of the error models used.

## III. Conclusion

We discussed a framework for error resilience in neuromorphic systems enabling reliable, low-power neuromorphic hardware. This will be critical for future scaling of AI systems. Acknowledgement: This research was supported by the U.S. National Science Foundation under Grant No. 2128149.

## References

[1] K. Boahen, "Dendrocentric learning for synthetic intelligence," Nature, vol. 612, no. 7938, pp. 43–50, Dec 2022. [Online]. Available: https://doi.org/10.1038/s41586-022-05340-6

[2] D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, and J. Dean, "Carbon emissions and large neural network training," arXiv preprint arXiv:2104.10350, 2021.

[3] M. M. S. Aly, T. F. Wu, A. Bartolo, Y. H. Malviya, W. Hwang, G. Hills, I. Markov, M. Wootters, M. M. Shulaker, H.-S. P. Wong et al., "The n3xt approach to energy-efficient abundant-data computing," Proceedings of the IEEE, vol. 107, no. 1, pp. 19–48, 2018.

[4] M. H. David Stutz, Nandhini Chandramoorthy and B. Schiele, "Bit error robustness for energy-efficient dnn accelerators," 2020.

[5] A. Saha, A. Chandramouli, and A. Chatterjee, "A resilience framework for synapse weight errors and firing threshold perturbations in rram spiking neural networks," Proceedings of the European Test Symposium (to appear), pp. 1–4, 2023.

[6] E. Ozen and A. Orailoglu, "Sanity-check: Boosting the reliability of safety-critical deep neural network applications," in 28th IEEE Asian Test Symposium, ATS 2019, Kolkata, India, December 10-13, 2019. IEEE, 2019, pp. 7–12. [Online]. Available: https://doi.org/10.1109/ATS47505.2019.000-8

[7] C. Amarnath, M. I. Momtaz, and A. Chatterjee, "Addressing soft error and security threats in dnns using learning driven algorithmic checks," in 2021 IEEE 27th International Symposium on On-Line Testing and Robust System Design (IOLTS), 2021, pp. 1–4.

[8] E. Ozen and A. Orailoglu, "Boosting bit-error resilience of dnn accelerators through median feature selection," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 39, pp. 1–1, 11 2020.

[9] Z. Chen, G. Li, and K. Pattabiraman, "A low-cost fault corrector for deep neural networks through range restriction," in 2021 51st Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN), 2021, pp. 1–13.

[10] C. Amarnath, M. Mejri, K. Ma, and A. Chatterjee, "Soft error resilient deep learning systems using neuron gradient statistics," in 2022 IEEE 28th International Symposium on On-Line Testing and Robust System Design (IOLTS). IEEE, 2022, pp. 1–7.

[11] K. Ma, A. Saha, C. Amarnath, and A. Chatterjee, "Efficient low cost alternative testing of analog crossbar arrays for deep neural networks," in 2022 IEEE International Test Conference (ITC). IEEE, 2022, pp. 499–503.