Out-of-Distribution Detection Using Power-Side Channels for Improving Functional Safety of Neural Network FPGA Accelerators

Vincent Meyers, Dennis Gnad, Mehdi Tahoori
Chair of Dependable Nano Computing (CDNC), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
{vincent.meyers, dennis.gnad, mehdi.tahoori}@kit.edu

***Abstract*—Accurate out-of-distribution (OOD) detection is crucial for ensuring the safety and reliability of neural network (NN) accelerators in real-world scenarios. This paper proposes a novel OOD detection approach for NN FPGA accelerators using remote power side-channel measurements. We assess different methods for distinguishing power measurements of in-distribution (ID) samples from OOD samples, comparing the effectiveness of simple power analysis and OOD sample identification based on the reconstruction error of an autoencoder (AE). Leveraging on-chip voltage sensors enables non-intrusive and concurrent remote OOD detection, eliminating the need for explicit labels or modifications to the underlying NN.**

***Index Terms*—out of distribution detection, neural network accelerators, power side channel**

I. Introduction

The rapid evolution of AI and deep learning has led to widespread neural network (NN) deployment in safety-critical applications like autonomous driving and medical diagnosis [1]. Specialized hardware accelerators, including GPUs, TPUs, and FPGAs, are essential for handling the growing complexity and size of NN models [2], enabling real-time analysis in various AI applications.

Despite their performance, NN accelerators face limitations, notably in detecting out-of-distribution (OOD) inputs, posing risks in safety-critical scenarios [1]. Detecting and handling OOD inputs are crucial for ensuring the reliability of NN accelerators. Various techniques, such as confidence thresholding, entropy-maximization and energy-based methods, aim to address this challenge [1], [3], [4].

Existing OOD-detection methods have limitations, often requiring computationally expensive procedures and reliance on input data distribution, introducing latency in real-time edge environments. We propose a novel OOD-detection approach using power side-channel measurements, leveraging on-chip FPGA voltage sensors. Our non-intrusive method extracts normal power patterns through simple power analysis (SPA) and an autoencoder (AE), enabling concurrent OOD detection during inference without model interference.

The contributions of this paper can be summarized as:

- Non-intrusive and concurrent power side-channel based detection of OOD samples for NN FPGA accelerators
- Evaluation and comparison of various methods for the detection: SPA and using AE reconstruction error
- An autoencoder (AE) architecture for learning a compressed representation of power traces of ID samples

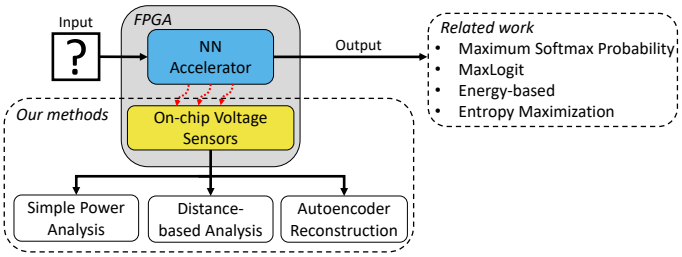


Fig. 1: Setup for our OOD detection and related work.

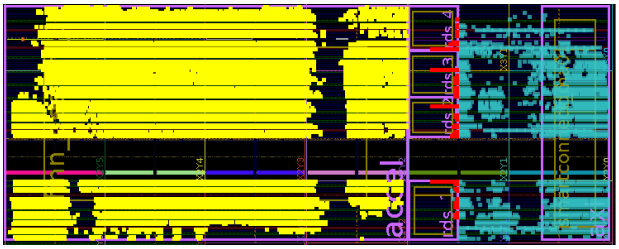


Fig. 2: Floorplan of the FINN accelerator (yellow) and multiple sensors (red), remaining control and debug logic (cyan).

II. Experimental Setup

We employ the publicly accessible FINN framework [5] to generate all neural network accelerators for our experiments. Subsequently, we deploy these accelerators onto a Zynq Ultrascale ZCU104. Routing delay sensors (RDS) [6] for measuring voltage fluctuations are co-located on the same FPGA as shown in Figure 1. An exemplary floorplan is displayed in Figure 2, with the accelerator in yellow and sensors in red.

III. Results

As a preliminary experiment, we compare the FPR-95 of the considered related work for different levels of quantization. The results for this experiment are presented in Figure 3, which shows the FPR-95 for the MLP64 with 1,2 and 4-Bit quantization. We notice that with increasing precision of weights and activations, the FPR-95 decreases, indicating an improvement of the detection mechanism. Therefore, in addition to the poor performance on the quantized model, we conclude that lowering the model's precision degrades the efficiency of existing methods.

Figure 4 (a) shows the ROC for SPA-based OOD detection. We notice that gaussian and uniform noise seem to have distinguishable features from the training set, while the detection is

979-8-3503-4859-0/DATE24/© 2024 EDAA

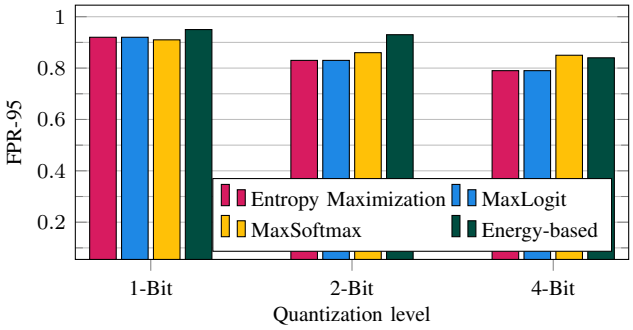


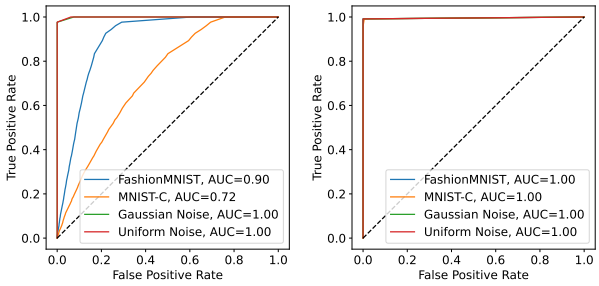
Fig. 3: Evaluation of FPR-95 for related work with three quantization levels and MNIST-C as the OOD dataset.

TABLE I: False positive rate at 95% TPR (FPR-95) and area under curve (AUC) of the OOD-Detection. ↓ means lower values are better and ↑ means high values are better.

Method	FashionMNIST		MNIST-C		Uniform & Gaussian Noise	
	FPR-95 ↓	AUC ↑	FPR-95 ↓	AUC ↑	FPR-95 ↓	AUC ↑
SPA	0.24	0.90	0.65	0.72	**0.00**	**1.00**
Euclidian Dist.	0.34	0.87	0.65	0.69	**0.00**	**1.00**
AE Recon. Error	**0.00**	**1.00**	**0.01**	**1.00**	**0.00**	**1.00**
Entropy Max. [4]	0.49	0.91	0.92	0.72	0.03	0.97
MaxLogit [7]	0.81	0.89	0.95	0.71	0.03	0.98
MaxSoftmax [1]	0.49	0.90	0.92	0.72	0.02	0.98
Energy-based [8]	0.73	0.88	0.91	0.7	0.05	0.95

much harder for FashionMNIST and MNIST-C. This is further stressed by the lower FPR95 score of the noise datasets. This method outperforms related work for MNIST-C and is also computationally efficient. It could be easily implemented on the FPGA itself as it requires no complex calculations.

Utilizing the AE trained on power traces of the training set, we perform the OOD detection on ID and OOD data. The threshold is set to the mean reconstruction error for the training set and then increased to reach a TPR of 95%. This is presented in Figure 4, which displays the performance of the AE trained on 60k traces (b). The AE method clearly outperforms all other methods with 0% FPR95 and 1.0 AUC. For all the evaluated datasets, we find that the hardest task is detection of samples from the MNIST-C set. These samples are closest to actual MNIST samples, making the detection a more challenging task. Still, we can achieve 100% OOD detection by employing an AE for reconstruction of the power traces.



(a) OOD detection with SPA (b) OOD detection with AE

Fig. 4: ROC of the OOD detection with SPA and AE

We find that existing methods [1], [4], [7], [8] perform similarly to our approach on random noise datasets. However, for datasets closer to the original MNIST range, their performance is inferior. Yang et al. [9] categorize these as near out-of-distribution (OOD) sets, with only a semantic shift from the ID dataset. Particularly for MNIST-C, we achieve a significantly lower FPR95, outperforming the best related work at 92% FPR95, while our worst method achieves 79% FPR95. Therefore, we conclude that our method surpasses existing methods in performance. The suboptimal performance of current methods in OOD detection may stem from their lack of adaptation to quantized neural networks (QNNs) with a reduced parameter space. The difference in QNN parameters could diminish the distinctiveness of logits and softmax values, reducing the effectiveness of existing methods. Our method remains effective even for the worst-case scenario of a binary NN, which exhibits a less pronounced power profile due to lower switching activity than less quantized networks.

IV. Conclusion

We propose an innovative approach to efficiently and effectively detect OOD samples in NN FPGA accelerators, using power side-channel measurements and the use of the reconstruction error of an AE model. In addition to using an AE, we evaluate detecting OOD based on SPA, which is a noteworthy second that can reduce computational effort. We show that both methods improve over the existing ones, which are not suitable for quantized NNs, as commonly used for edge hardware deployment. The proposed approaches enhance the robustness and reliability of edge accelerators in safety-critical domains, ensuring their safe and secure operation even in the presence of OOD inputs. Additionally, our method can be extended to different domains and datasets, enabling the detection of OOD inputs in a wide range of application scenarios. In the future, we are looking into expanding our experiments and method for other types of NNs, such as CNNs, and tasks beyond image classification.

References

[1] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv preprint arXiv:1610.02136*, 2016.

[2] Microsoft. (2022, Apr.) Deploy ML models to field-programmable gate arrays (FPGAs) with Azure Machine Learning. [Online]. Available: https://learn.microsoft.com/en-us/azure/machine-learning/v1/how-to-deploy-fpga-web-service

[3] T. DeVries and G. W. Taylor, "Learning confidence for out-of-distribution detection in neural networks," *arXiv preprint arXiv:1802.04865*, 2018.

[4] R. Chan *et al.*, "Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation," in *International Conference on Computer Vision*. IEEE, 2021.

[5] Y. Umuroglu *et al.*, "FINN: A Framework for Fast, Scalable Binarized Neural Network Inference," 12 2016.

[6] D. Spielmann *et al.*, "RDS: FPGA routing delay sensors for effective remote power analysis attacks," *IACR Transactions on Cryptographic Hardware and Embedded Systems*, 2023.

[7] D. Hendrycks *et al.*, "Scaling out-of-distribution detection for real-world settings," *arXiv preprint arXiv:1911.11132*, 2019.

[8] W. Liu *et al.*, "Energy-based out-of-distribution detection," *Advances in Neural Information Processing Systems*, 2020.

[9] J. Yang *et al.*, "Openood: Benchmarking generalized out-of-distribution detection," *Advances in Neural Information Processing Systems*, 2022.