Harnessing ML Privacy by Design Through Crossbar Array Non-idealities

Md Shohidul Islam*§, Sankha B. Dutta §, Andres Marquez§, Ihsen Alouani‡, Khaled N. Khasawneh*
*ECE Dept., George Mason University, Fairfax, VA, USA
§Pacific Northwest National Lab, Richland, WA, USA
‡CSIT, Queen's University Belfast, UK
Email: {mislam20, kkhasawn}@gmu.edu, {sankha.b.dutta, andres.marquez}@pnnl.gov, i.alouani@qub.ac.uk

***Abstract*—Deep Neural Networks (DNNs), handling compute- and data-intensive tasks, often utilize accelerators like Resistive-switching Random-access Memory (RRAM) crossbar for energy-efficient in-memory computation. Despite RRAM's inherent non-idealities causing deviations in DNN output, this study transforms the weakness into strength. By leveraging RRAM non-idealities, the research enhances privacy protection against Membership Inference Attacks (MIAs), which reveal private information from training data. RRAM non-idealities disrupt MIA features, increasing model robustness and revealing a privacy-accuracy tradeoff. Empirical results with four MIAs and DNNs trained on different datasets demonstrate significant privacy leakage reduction with a minor accuracy drop (*e.g.*, up to 2.8% for ResNet-18 with CIFAR-100).**

I. Introduction

Deep Learning, widely applied across diverse fields, relies on powerful Deep Neural Networks (DNNs) and Convolutional Neural Networks (CNNs). Specialized accelerators address the compute-intensive Multiplication-and-Accumulation (MAC) computations inherent in DNNs. Yet, the limitations of digital CMOS hardware have spurred interest in Non-Volatile Memory (NVM)-based accelerators such as RRAM [1]. RRAM, organized as crossbars, enables direct MAC computations within the memory array, reducing data movement and yielding significant time and energy savings in DNN computation.

In our study, we empirically investigate RRAM crossbar arrays, known for analog computation. These crossbars have inherent non-idealities from the device, access transistors, and parasitic resistors, causing outputs to deviate from ideal values. This approximation poses reliability challenges for general applications. Interestingly, DNNs/CNNs, being highly fault-tolerant, can benefit from in-situ computing with RRAM crossbars for acceleration. Recent research suggests that, in addition to accelerating DNNs, RRAM crossbars also contribute to the robustness of DNNs against adversarial examples.

The study investigates whether the non-idealities of RRAM crossbar can enhance the robustness of Deep Neural Networks (DNNs) against privacy attacks, particularly Membership Inference Attacks (MIAs) [2]. MIAs can leake sensitive private information of a model through inference. Given the increasing use of sensitive datasets (*e.g.,* genome data, personal photos, clinical/biomedical records, location data, etc) and the training of machine learning models on cloud platforms, data privacy, especially in the context of trustworthy machine learning systems, is a critical concern.

To enhance privacy in machine learning models, existing defenses like differential privacy and output obfuscation compromise utility and demand retraining. Privacy Preserving Voltage (V_{PP} [3]) provides a promising tradeoff but requires hardware arrangements and is device-specific. In contrast, RRAM crossbar, explored in this study, inherently disrupts privacy attack features, offering privacy gains to DNNs without extra computation or device-specific constraints. This research, using the PytorX framework, marks the first exploration of ML privacy through RRAM crossbar implementation.

II. Experimental Settings

This section details the experimental configurations, covering DNN models, datasets, MIA attacks, and RRAM crossbar dimensions. In line with previous studies, we utilize four common datasets: Purchase100 (with fully connected DNN), Texas100 (with fully connected DNN), CIFAR10 (with AlexNet), and CIFAR100 (with ResNet18). Privacy and utility are assessed by implementing target DNNs on RRAM crossbar with varying dimensions (32×32, 64×64, and 128×128). Privacy risk is measured against three MIA attacks (I_{bb} [4], I_{bl} [5], and I_{nn} [6]), each using distinct features for inference— 'confidence and loss' for I_{bb}, 'logit' for I_{nn}, and 'loss' for I_{bl}.

III. Evaluation

A. Utility Evaluation

"Utility" refers to the classification accuracy of target models. For any model and application, it is always expected to uphold a high level of utility even when defense mechanisms are in place. However, the presence of non-idealities in RRAM crossbars contributes to perturbations in model computations, which is expected to impact the classification accuracy.

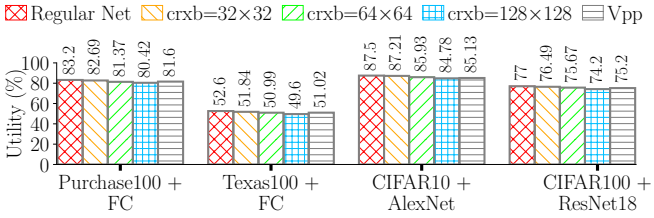

Fig. 1: Baseline accuracy of different models.

979-8-3503-4859-0/DATE24/© 2024 EDAA

In Figure 1, we present the utility of Regular Net (*i.e.*, model that is not implemented on RRAM crossbar) without applying any defense, RRAM crossbar net, and Regular Net under the most effective existing V_{PP} defense. When compared to the Regular Net, crossbar models exhibit a reduction in accuracy for all crossbar sizes. Considering the example of ResNet18 with CIFAR100 dataset, we observe a 0.51%, 1.33%, and 2.8% utility drop for RRAM crossbar networks with crossbar dimension of 32×32, 64×64, and 128×128, respectively. Besides ResNet18, we observe the similar trend of utility drop for other crossbar DNNs and datasets as well, *i.e.*, Purchase100+FC, Texas100+FC, and CIFAR10+AlexNet.

B. Privacy Evaluation

In this section, we delve into the assessment of privacy risk associated with both unprotected regular net and crossbar based target models. Additionally, we conduct a comparative analysis of the privacy risk stemming from crossbar models against that of existing defenses. In line with established methodologies, we quantify privacy risk using MIA accuracy. Note that the privacy risk is at its minimum when attackers are most uncertain, typically occurring when their prediction resembles a random guess or when the MIA accuracy stands at 50%. Conversely, an MIA accuracy surpassing 50% implies a heightened level of privacy risk.

TABLE I: Privacy risk of crossbar based models, undefended regular models, and V_{PP} [3]-protected models.

Dataset + model	Defense	MIA accuracy (privacy risk)		
		$I_{bb}(\%)$	$I_{bl}(\%)$	$I_{nn}(\%)$
Purchase100 + FC	None	77.09	63.6	62.2
	$crxb = 32 \times 32$	72.52	61.08	60.32
	$crxb = 64 \times 64$	66.18	59.33	58.81
	$crxb = 128 \times 128$	59.70	57.42	56.44
	V_{PP} [3]	53.27	53.20	50.16
Texas100 + FC	None	76.23	76.2	72.0
	$crxb = 32 \times 32$	71.19	73.66	67.05
	$crxb = 64 \times 64$	65.03	67.18	62.71
	$crxb = 128 \times 128$	57.41	60.82	58.09
	V_{PP} [3]	53.60	52.40	50.07
CIFAR10 + AlexNet	None	78.1	66.96	66.15
	$crxb = 32 \times 32$	72.06	65.19	65.79
	$crxb = 64 \times 64$	68.22	62.13	63.06
	$crxb = 128 \times 128$	61.36	59.64	60.18
	V_{PP} [3]	51.94	51.85	50.83
CIFAR100 + ResNet18	None	77.49	68.3	67.1
	$crxb = 32 \times 32$	74.65	66.73	65.42
	$crxb = 64 \times 64$	69.04	63.17	63.05
	$crxb = 128 \times 128$	60.75	58.82	59.93
	V_{PP} [3]	53.38	53.37	50.64

Table I provides an overview of MIA accuracy, which represents the privacy risk, against the score based attacks I_{bb}, I_{bl}, and I_{nn}. In the 'Defense' column, 'None' refers to the unprotected regular net and $crxb$ indicates crossbar models. The table clearly illustrates that unprotected models exhibit substantial privacy risks, whereas crossbar models exhibit a noteworthy reduction in privacy risks. Moreover, models with higher crossbar dimension preserves higher privacy. To investigate the reason, we analyze the confidence distribution for training (member) and testing (non-member) samples. The 'regular nets' exhibit minimal overlap between the scores of members and non-members, indicating a lower level of confusion and better distinguishability. Conversely, the crossbar models display more significant overlaps, signifying increased confusion and reduced separability. Furthermore, increasing the crossbar dimension causes greater overlaps/confusion for the MIA attack algorithm, which in turn preserves higher privacy. Another finding is that, while crossbar models preserves better privacy than regular unprotected models, existing V_{PP} protected models outperforms across all attacks. Note that crossbar models offer by-product privacy gain while V_{PP} requires additional hardware/software for undervolting, with the risk of system crash and device-specific deployment.

IV. Conclusion

This paper unprecedentedly explores the by-product privacy harnessing of ML models through RRAM crossbar array implementation. Empirical results demonstrate that RRAM non-idealitites can garble the privacy attack features and thus weakens the attack success. Investigation using four state-of-the-art privacy attacks shows that RRAM based models significantly reduce privacy leakage (*e.g.*, upto 25.86%) as compared to regular non-crossbar based DNNs/CNNs, with a nominal accuracy loss (*e.g.*, a maximum of 2.8%).

Acknowledgment

This work was supported by the U.S. DOE Office of Science, Office of Advanced Scientific Computing Research, under awards 66150: "CENATE - Center for Advanced Architecture Evaluation" and 76125: "AMAIS - Advanced Memory to support Artificial Intelligence for Science" . The Pacific Northwest National Laboratory is operated by Battelle for the U.S. Department of Energy under contract DE-AC05-76RL01830. The work was also partially supported by US National Science Foundation grants CNS-1955650, CNS-2053383, and CCF-2212427, as well as EdgeAI KDT-JU European project (101097300).

References

[1] H.-S.P. Wong *et al.*, "Metal–oxide rram," *Proceedings of the IEEE*, vol. 100, no. 6, pp. 1951–1970, 2012.

[2] R. Shokri *et al.*, "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017, pp. 3–18.

[3] M.S. Islam *et al.*, "Vpp: Privacy preserving machine learning via undervolting," in *2023 IEEE International Symposium on Hardware Oriented Security and Trust (HOST)*. IEEE, 2023, pp. 315–325.

[4] M. Nasr *et al.*, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *2019 IEEE symposium on security and privacy (SP)*. IEEE, 2019, pp. 739–753.

[5] S. Yeom *et al.*, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*. IEEE, 2018, pp. 268–282.

[6] A. Salem *et al.*, "Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models," in *Proceedings of the 2019 Network and Distributed System Security Symposium (NDSS)*, 2019.