A Privacy-aware Quantilisation Approach for Efficient Edge Deep Learning Accelerator

Huaizhi Zhang, Xuqi Zhu, Jiacheng Zhu, Klaus McDonald-Maier and Xiaojun Zhai
University of Essex, Colchester, United Kingdom
{hz24245, xz18173, jz23222, kdm, xzhai}@essex.ac.uk

***Abstract*—Data privacy is one of the key concerns in machine learning model applications at the edge, especially in sensitive domains such as the healthcare sector. Here adversaries may exploit Membership Inference Attacks (MIAs) to determine if particular data points were used as part of datasets from the model's training set, potentially leading to further data leakage issues. Although privacy preservation techniques like differential privacy (DP) can mitigate such risks during the training phase, this often results in degradation of model accuracy, making them less suitable for cloud training and edge deployment paradigms. For AI edge applications, existing research works for designing neural network accelerators primarily prioritize computational performance and power efficiency as their design target. In this paper, we introduce a novel privacy-aware quantilisation approach for deep learning accelerators and analyse the trade-off between computational efficiency and privacy protection. The proposed system allows to adjust privacy constraints through tunable parameters, enabling flexible deployment on edge devices while meeting privacy and performance constraints. We have evaluated the proposed design on an AMD VCK190 board using a range of hypothetical MIA benchmarks. The results demonstrate that the proposed approach can effectively reduce the success rate of MIA attacks across multiple performance metrics.**

***Index Terms*—Membership Inference Attack (MIA), Quantisation, Privacy Protection, Accelerator**

I. Introduction

With substantially growing number of edge devices, machine learning (ML) algorithms are increasingly being deployed directly on these edge platforms, thus enabling real-time data processing and ultra-low latency by minimising the need to send data to centralized servers for computation. However, compared to the centralised server based inference solutions, such edge devices are more vulnerable to cyber attacks. Server-based services are typically hosted within centralised cloud infrastructures, which benefit from advanced management and security mechanisms, including sophisticated firewalls, comprehensive monitoring systems, and stringent access controls. These measures significantly enhance the server's resilience against adversarial attacks. In contrast, edge devices typically have limited computational resources and weaker cyber security mechanisms, making them more vulnerable to attacks. This vulnerability is especially concerning in sectors like industry, healthcare, and finance, where data privacy is paramount.

Due to strict resource and power limitations of the edge devices, much of the prior research has focused on developing dedicated accelerator architectures and compressed ML algorithms to optimise throughput and energy efficiency whilst minimising hardware footprints across various applications. For example, these papers [1], [2] focus on designing hardware architectures for optimising General Matrix Multiplication (GEMM). In parallel, FP-BNN [3] and FINN [4] present binary quantisation techniques to reduce the computation and data movements overhead. These works have achieved impressive improvements in terms of Power consumption, Performance, and Area occupancy (PPA). However, these works have largely ignored the privacy needs of edge devices, despite the growing importance of data protection. Shokri et al. [5] suggest that machine learning algorithms have exposed the vulnerability in leaking sensitive information from the training dataset by detecting whether a specific record was involved in the training dataset of the target model. This attack, known as Membership Inference Attack (MIA), exacerbates data privacy concerns, particularly for edge devices with limited monitoring and regulatory capabilities.

To mitigate the risks of MIA, researchers in [6], [7] used Differential Privacy (DP) to effectively defend against MIA. Additionally, Jia et al. [8] introduced MemGuard, the first defense mechanism to offer formal guarantees on utility loss against black-box MIA. However, these solutions only concentrate on algorithm design, overlooking the importance of hardware design as well as the hardware/software co-design in ensuring data privacy on edge devices. Therefore, these approaches have limited supports for deploying valid defense mechanisms in resource-constrained edge environments, or provide insufficient performance due to running computation-intensive algorithms beyond the capabilities of edge devices.

To execute privacy-aware ML application on edge devices, this paper proposes a versatile FPGA-based accelerator integrated with privacy-aware model quantilisation algorithms by leveraging hardware/software co-design approaches. The main contributions of this work can be summarised as follows:

1) We investigate the impact of model quantisation on model privacy, and propose a privacy-aware quantisation scheme that can significantly improve model privacy.

2) We design a scalable, parameterised edge deep learning accelerator capable of optimising the trade-off between computational performance and privacy preservation.

3) We evaluate multiple MIA attack scenarios on a range of models, and experimental results show that the proposed system is able to achieve up to 23% reduction in the attack success rate.

The remainder of the paper is organised as follows: Section

II covers membership inference attack models, metrics, and ML quantification schemes. Section III details the design of our privacy-preserving system, and Section IV evaluates it using multiple MIA benchmarks. Section V concludes the paper.

II. Related work

MIA refers to an adversary inferring whether a specific data point was used in the model's training process by querying the model. Shokri et al. [5] were the first to propose the attack, designing a shadow model to simulate the target model and successfully implement the attack using synthetic data. They also analysed the attack's effectiveness, and their attack model is illustrated below in Fig. 1.

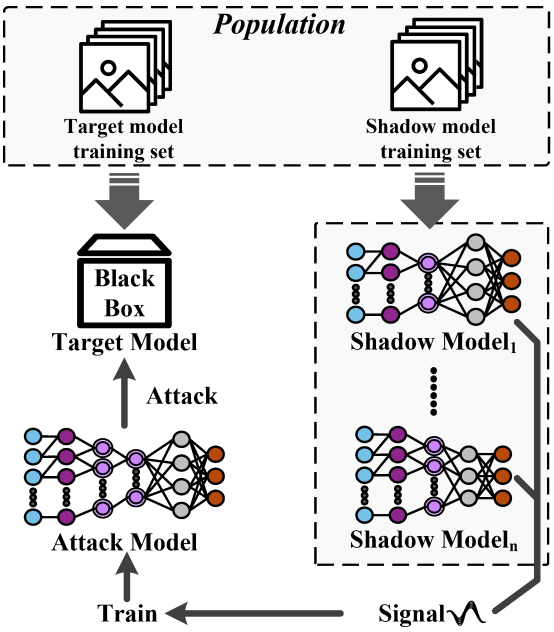

Fig. 1: An example of shadow attack in black box setting

In this attack pattern the adversary distinguishes whether a data point belongs to the training set of the model by comparing the predicted responses of the shadow model and the target model for the same data point. Yeom et al. [9] conducted a more intensive analysis of why models are successfully attacked and found that the further the model is overfitted, the more it is vulnerable to attack. Subsequently, Shokri et al [10], [11] expanded on their work by developing a series of more advanced attacks based on varying levels of adversarial prior knowledge. They also introduced the ML Privacy Meter tool, which enables quantitative analysis of privacy risks in ML systems and aids in conducting privacy audits [12]. LiRA is a likelihood ratio-based type of MIA proposed by Cam et al [13]. It identifies members and non-members by analysing the distributional properties of model outputs, and achieves a significant attack success rate with low false positive rates. Recently, the AWS research team highlighted that previous attacks heavily relied on shadow models, which introduce significant computational overhead. To address this, they proposed a membership inference attack based on quantile regression [14], which only requires a single shadow model. This approach achieves higher attack accuracy than LiRA, while requiring zero prior knowledge of the target model.

III. Methodology

The proposed system overview is shown in Fig. 2 for MIA resistance, which contains the privacy aware quantise algorithm and the corresponding FPGA based hardware system capable of deploying the quantised models.

A. Privacy-aware Quantization

As the underlying idea of MIA is the inevitable overfitting of training data in the model, so the distribution of member's signals (i.e the model loss) are moderately different from that of Non-member's. Therefore the goal of the proposed privacy-aware quantisation approach is to constrain the difference between the distributions of members and non-members on the inference output of the quantised model f, i.e., Eq 1.

$$\arg\min_{f\in q} \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}|f(x_i^{\text{mem}}) - f(x_i^{\text{non}})| \,\middle|\, \text{label} = y\right] \quad (1)$$

To determine differences in model signals, we iterate over all data in the same categories and measure the distribution of activation values in each layer and count the differences between members and non-members. The statistical results are shown in the Fig. 4(a), for the same classification inputs, the distribution error of the model's activation values in the model's feature extraction phase is relatively large, but as the model layer gets deeper, their distribution difference will start to decline. Quantisation in deeper layers of the model therefore reduces the difference between members and non-members, which consequently decreases the privacy breaches. In order to formulate the consequences of quantisation, we analyse the quantisation process and the impact of quantisation on member and non-member variance. Given the weight matrix $\mathbf{W}$, the linear layer can be written as: $y = \mathbf{W}x$. The quantised version of this linear layer can be written as: $y_q = \mathbf{W}_q x_q$. The quantisation process is defined as:

$$s_W = \frac{W_{\text{max}} - W_{\text{min}}}{2^b - 1}, \quad z_W = \left\lfloor \frac{W_{\text{min}}}{s_W} \right\rceil \quad (2)$$

$$W_q = \text{round}\left(\frac{W}{s_W} + z_W\right) \quad (3)$$

where b is the quantisation bit width, s_W is the scaling factor. The round operation in the quantisation process introduces an error denoted as $RoundErr$, which usually leads to a decrease in model accuracy. However, statistically we are surprised to discover that it is able to decrease the variance between member and non-member. As shown in Fig 4(b), depending on the value of the member, the $RoundErr$ decreases the variance between member's activation value and the non-members' with varying probability (e.g., 70% shown in the Fig 4(b). Further, we statistically calculate the difference between the activation values of member and non-member, as shown

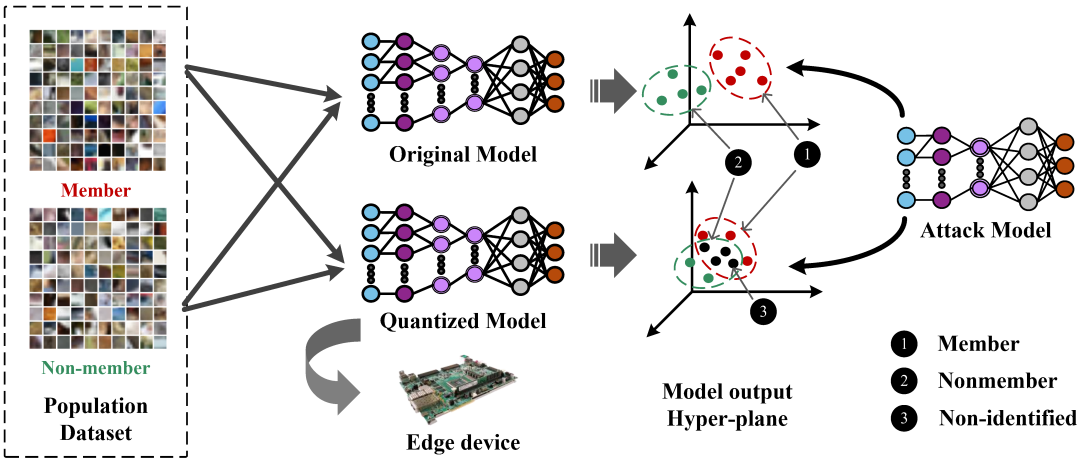

Fig. 2: Overview of the design methodology

in Fig 4(c), and we are able to eliminate more than 95% of the difference using an appropriate bit width. In order to maintain the accuracy of the model, we only perform low-bit quantisation in specific layers and maintain higher accuracy quantisation in other layers, thus enabling a smaller accuracy degradation for the entire model. Using a range of MIA attacks as evaluation benchmark, we verify that this approach is effective in enhancing the privacy performance of the model, whilst maintaining the target model's accuracy as shown in Section IV.

B. Hardware Architecture

Since the proposed solution uses different quantisation bit widths in a model, it requires a more sophisticated computational platform to achieve high computational efficiency. We have thus designed a multi-precision neural network accelerator, using AMD's Deep Learning Processor Unit (DPUs) [15], and its customised hardware architecture is shown in Fig 3.

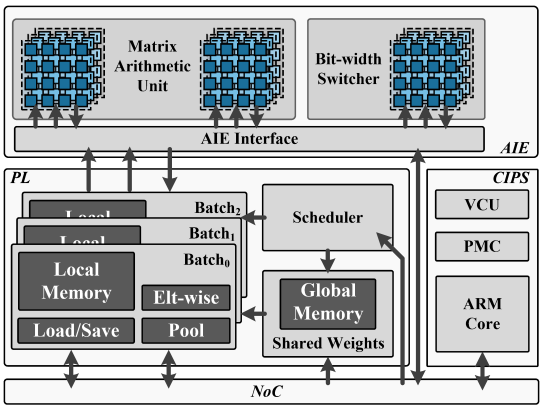

Fig. 3: Customised DPU hardware architecture

In this custom deep learning accelerator, we introduce an additional bit-width switching module, where we use re-linearisation instead of direct bit-width interception algorithm to achieve better privacy performance. By re-quantising the module's input, it amplifies the contribution of $RoundErr$, which improves the privacy performance. For computational performance, since deep neural network inference is a memory-intensive task, reducing data migration can significantly reduce system power consumption. In the proposed design, we use the Zero-copy technique [16] to achieve a significant reduction in system power consumption by eliminating the need for multiple copy operations between the ARM core and the FPGA, when transferring data between different modules.

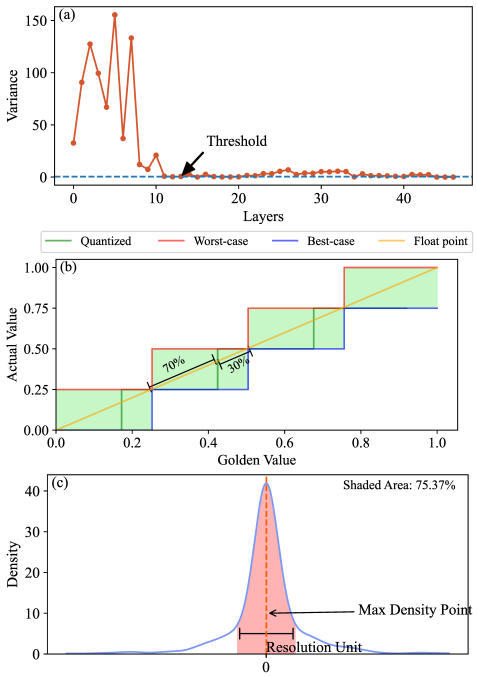

Fig. 4: Principle of privacy awareness quantitative operations for member and non-member: (a) behaviour differences across the model, (b) impact of $RoundErr$ on their differences, (c) distribution variance

IV. Evaluation

We evaluate the proposed system on image classification benchmarks, but it is worth noting that the proposed approach

is not specific to a particular model and it can therefore work for other deep learning tasks too.

A. Target Model

We train all the target model on the CIFAR-10 and CIFAR-100 datasets [17]. In order to verify the generality of the proposed system, we train models such as vgg19 [18] and resnet as the target model respectively, and also according to previous studies, the training algorithm does affect the privacy performance, thus we conduct the control variable experiments and use the same Stochastic Gradient Descent(SGD) algorithm for model training.

B. Threat Models

In this paper, we quantitatively analyse the privacy preservation performance of the proposed system using the accuracy and AUC of different MIA attacks as metrics. Adversaries conducting LiRA and Reference attacks will follow these assumptions [5]: (1) The reference dataset $\mathcal{D}_{reference}$ (i.e. $\mathcal{D}_{reference} \overset{\text{i.i.d.}}{\underset{s'}{\sim}} \mathcal{D}_{public}$) is available to the adversary, along with the training set $\mathcal{D}_{private}$ (i.e. $\mathcal{D}_{private} \overset{\text{i.i.d.}}{\underset{s}{\sim}} \mathcal{D}_{public}$) of the target model, are both randomly sampled from the same underlying data distribution $\mathcal{D}_{public}$. (2) The adversary has access to the structure of the target model, but not to the weights and activation values.

For the reference attack, we train 16 reference models on different reference datasets separately. We employ the quantile of the model loss distribution to determine a threshold to split members and non-members, with the threshold determined by the target false positive rate. For LiRA, similar to the reference attack, we firstly train the shadow models on $\mathcal{D}_{reference}$, we use 16 shadow models for calculating the score based on the model loss, and determine the distribution of members and non-members according to the scores of the training dataset and testing dataset, and finally we determine the probability if input data is in the training set of the target model using Eq.4, where q is the target model, μ and σ are the mean and standard deviation of members' scores, respectively.

$$\log f(\text{loss}(x)_q \mid x \in D_{\text{Train}}) = -\frac{1}{2}\left(\log(2\pi) + \log(\sigma^2) + \frac{(x-\mu)^2}{\sigma^2}\right) \quad (4)$$

C. Performance evaluation

We conduct evaluations experiments using Reference attack and LiRA as benchmarks, respectively. For VGG [18], whose variance shown in Fig 4(a), we choose the first layer (layer 13) with variance between member and non-member that are lower than the threshold (i.e. hyper-parameter) to apply the quantisation. After conducting experiments on ResNet [19] and DenseNet [20] in the same manner, we obtain the results shown in Table I. The proposed system reduces the success rate of attacks by 9% to 24% on different benchmarks (e.g. LiRA on ResNet and reference attack on ResNet) and demonstrates the ability to more effectively protect the data privacy of the training set.

D. Trade-off analysis

Although the proposed quantisation algorithms can enhance privacy preservation performance, they may lead to a decrease in model accuracy. However, quantisation also significantly reduces the model size, which in turn lowers the power consumption during deployment. Therefore, a careful trade-off analysis should be considered in this multi-objective optimisation task. We conduct a series of experiments to demonstrate the impact of varying quantisation bit widths, providing a reference for different model application scenarios. As shown in Fig.5, the accuracy of the model decreases with decreasing quantisation bit-width, but not linearly, with a large degradation in accuracy occurring at 4-bit and lower quantisation bit-widths, but retaining high accuracy at bit-widths above 8-bit. As for the resistance to MIA, lower bit-widths show more resistance, especially when the bit-width is lower than 8-bit, as shown in Fig. 4(c), the $RoundErr$ almost leads to the variance between the member and non-member's views to zero, and therefore the accuracy of MIA attacks shows a significant degradation.

TABLE I: Performance Comparison on benchmarks

		LiRA		Reference Attack	
		AUC	Accuracy	AUC	Accuracy
VGG19	FP32	0.77	0.73	0.56	0.55
	Quantised	0.61	0.60	0.51	0.51
ResNet50	FP32	0.83	0.80	0.71	0.69
	Quantised	0.75	0.68	0.54	0.53
DenseNet121	FP32	0.81	0.77	0.71	0.70
	Quantised	0.72	0.69	0.59	0.57

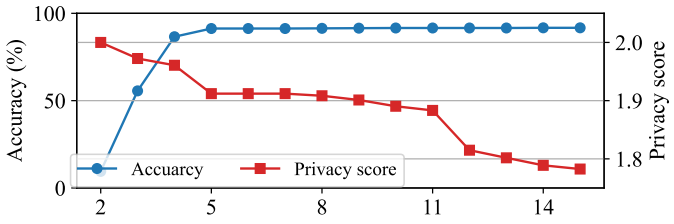

Fig. 5: Trade off analysis for model accuracy and privacy

V. Conclusion

This paper addresses the critical concern of data privacy in edge deep learning applications, particularly in sensitive sectors like healthcare. Edge devices are more vulnerable to MIA compared to centralised approaches. This paper introduces a system that balances privacy preservation and computing efficiency performances on edge devices through a privacy-aware algorithm, utilising a hardware-software co-design methodology. The proposed system's effectiveness is evaluated using LiRA and reference attacks on an AMD VCK190 board, demonstrating a significant reduction in the success rate of such MIA attacks.

ACKNOWLEDGMENT

This work is supported by the UK Engineering and Physical Sciences Research Council through grants, EP/X015955/1, EP/X019160/1 and EP/V000462/1, EP/V034111/1. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

REFERENCES

[1] T. Adiono, A. Putra, N. Sutisna, I. Syafalni, and R. Mulyawan, "Low latency yolov3-tiny accelerator for low-cost fpga using general matrix multiplication principle," *IEEE Access*, vol. 9, pp. 141 890–141 913, 2021.

[2] A. Ahmad and M. A. Pasha, "Optimizing hardware accelerated general matrix-matrix multiplication for cnns on fpgas," *IEEE TCS II: Express Briefs*, vol. 67, pp. 2692–2696, 2020.

[3] S. Liang, S. Yin, L. Liu, W. Luk, and S. Wei, "Fp-bnn: Binarized neural network on fpga," *Neurocomputing*, vol. 275, pp. 1072–1086, 1 2018.

[4] Y. Umuroglu and N. J. Fraser, "Finn: A framework for fast, scalable binarized neural network inference," *FPGA 2017*, pp. 65–74, 2017.

[5] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE symposium on security and privacy (SP)*. IEEE, 2017, pp. 3–18.

[6] J. Chen, W. H. Wang, and X. Shi, "Differential privacy protection against membership inference attack on machine learning for genomic data," in *BIOCOMPUTING 2021: Proceedings of the Pacific Symposium*. World Scientific, 2020, pp. 26–37.

[7] S. Rahimian, T. Orekondy, and M. Fritz, "Differential privacy defenses and sampling attacks for membership inference," in *Proceedings of the 14th ACM workshop on artificial intelligence and security*, 2021, pp. 193–202.

[8] J. Jia, A. Salem, M. Backes, Y. Zhang, and N. Z. Gong, "Memguard: Defending against black-box membership inference attacks via adversarial examples," in *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, 2019, pp. 259–274.

[9] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, 2018, pp. 268–282.

[10] J. Ye, A. Maddi, S. K. Murakonda, V. Bindschaedler, and R. Shokri, "Enhanced membership inference attacks against machine learning models," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 3093–3106. [Online]. Available: https://doi.org/10.1145/3548606.3560675

[11] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning," in *Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP)*, vol. 2018, 2018, pp. 1–15.

[12] S. K. Murakonda and R. Shokri, "Ml privacy meter: Aiding regulatory compliance by quantifying the privacy risks of machine learning," *arXiv preprint arXiv:2007.09339*, 2020.

[13] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer, "Membership inference attacks from first principles," in *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022, pp. 1897–1914.

[14] M. Bertran, S. Tang, A. Roth, M. Kearns, J. H. Morgenstern, and S. Z. Wu, "Scalable membership inference attacks via quantile regression," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[15] *The AMD Xilinx® Versal® Deep Learning Processing Unit (DPUCVDX8G)*, Advanced Micro Devices, Inc., 2023. [Online]. Available: https://docs.amd.com/r/en-US/pg425-dpucv2dx8g/

[16] O. Bell, C. Gill, and X. Zhang, "Hardware acceleration with zero-copy memory management for heterogeneous computing," in *2023 IEEE 29th International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA)*, 2023, pp. 28–37.

[17] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[20] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.