Quantization Methods of Defense against Membership Inference Attacks

Subjects: Computer Science, Artificial Intelligence

Contributor: Azadeh Famili , Yingjie Lao

Machine learning deployment on edge devices has faced challenges such as computational costs and privacy issues. Membership inference attack (MIA) refers to the attack where the adversary aims to infer whether a data sample belongs to the training set. In other words, user data privacy might be compromised by MIA from a well-trained model. Therefore, it is vital to have defense mechanisms in place to protect training data, especially in privacy-sensitive applications such as healthcare.

membership inference attack defense MIA

1. Introduction

Machine learning is an evolving field that has recently gained significant attention and importance. With the exponential growth of data and advancements in computing power, machine learning has become a powerful tool for extracting valuable insights, making predictions, and automating complex tasks. Significant advancements in machine learning have led to the remarkable performance of neural networks in a wide range of tasks [1][2]. As the demand for real-time processing and low-latency applications continues to rise, the importance of efficient hardware implementations of machine learning algorithms becomes evident. Hardware acceleration plays a crucial role in meeting the computational requirements and enabling the deployment of machine learning models in resource-constrained environments.

To facilitate the efficient deployment of machine learning models on hardware platforms, scientists and researchers have proposed compression techniques to accelerate training and inference processes. To this end, one of the promising techniques in model compression is quantization. Quantization methods [3][4][5] accelerate the computation by executing the operations with reduced precision. These methodologies have achieved performance levels comparable to those of full bitwidth networks while remaining compatible with resource-constrained devices. These methods also enable broader possibilities for machine learning applications, particularly in sectors that handle sensitive data on the edge.

This approach also proves valuable in various use cases, such as medical imaging [6], autonomous driving [7], facial recognition [8], and natural language processing [2], where the data privacy is of utmost importance. However, as these technologies become increasingly intertwined with daily life, they must be continuously evaluated for vulnerabilities and privacy concerns. For example, as shown in **Figure 1**, patient data can be used to train neural networks. In most cases, hospitals or healthcare providers gather a large amount of data regarding patients'

identity, health, insurance, and finance information. An adversary may attempt to gain access to this information at every step of this process, compromising user data privacy in machine learning applications.

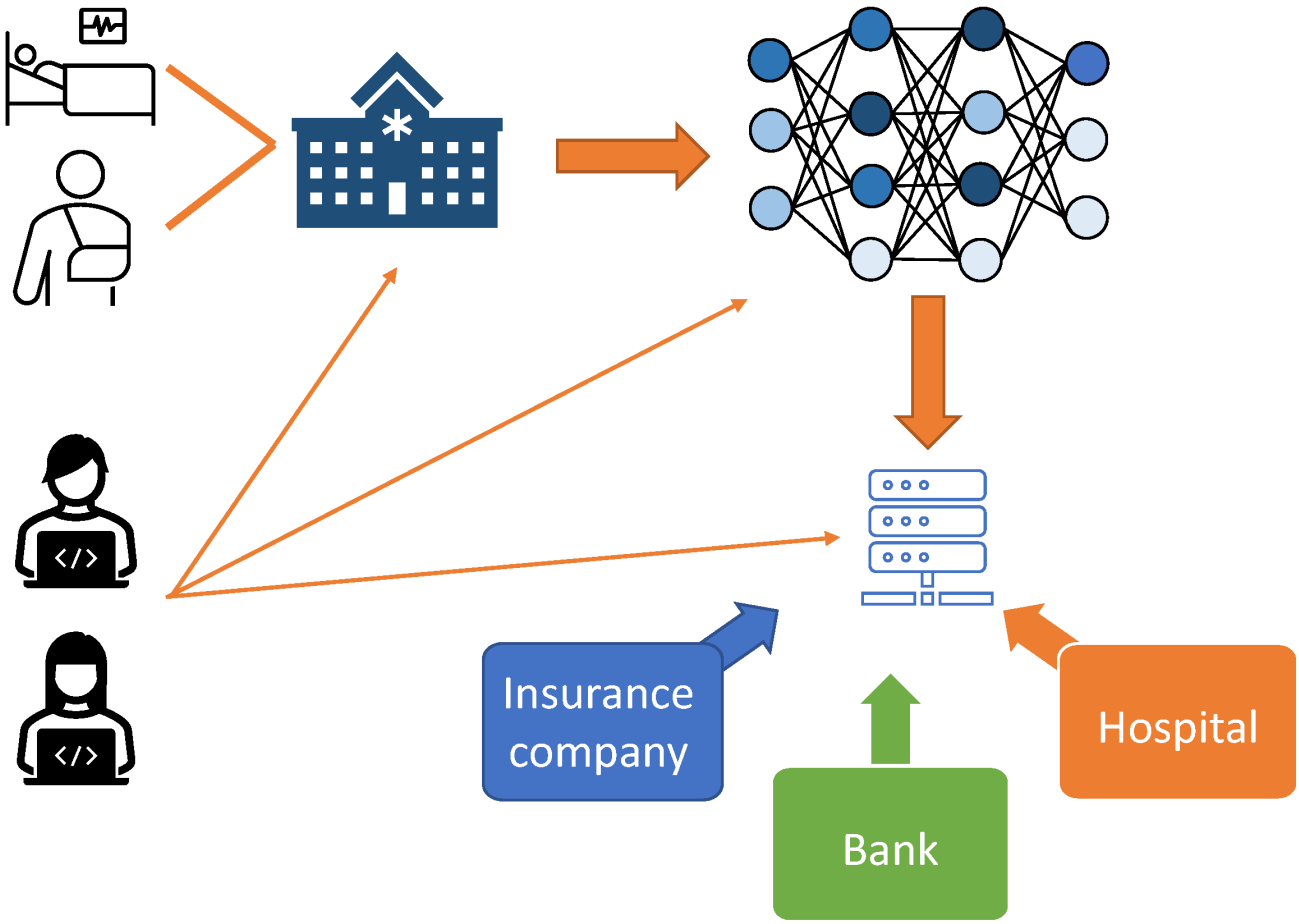

Figure 1. A patient medical and personal information is valuable in the field of machine learning. An adversary can jeopardize patient privacy from the machine learning models that are trained on the data.

2. Quantization Methods of Defense against Membership Inference Attacks

The issue of privacy attacks in neural network training applications has raised significant concerns, particularly in sensitive scenarios [9]. Extensive research has been conducted to address the privacy implications associated with training data, focusing on various aspects such as data leakage, prevention of memorization, and evaluation of the privacy efficacy of proposed defense mechanisms. Among these, MIA has emerged as a critical concern to user data privacy in machine learning applications, as it has been shown that MIA can effectively determine whether a data sample belongs to the training set. Such MIA methods are able to extract the user data information contained in the overparameterized model. The high-level overview of MIA is shown in **Figure 2**. An adversary passes a data sample *x* to the target model using some analysis tools to determine the membership of this data sample.

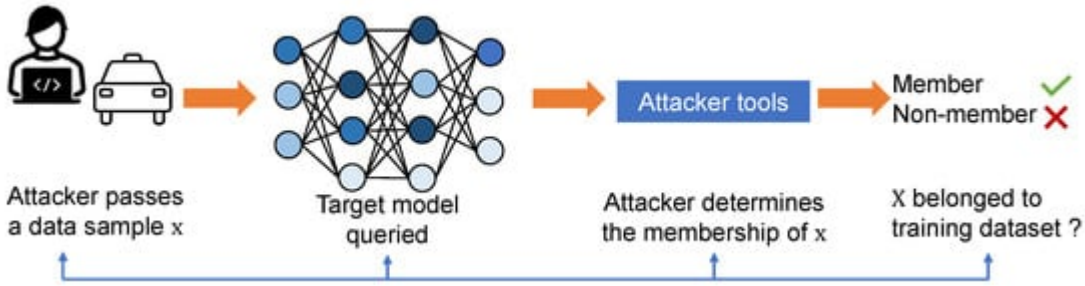

Figure 2. Overview of MIA attack. Here x is a data sample which the attacker wants to determine its membership.

The first MIA approach [10] uses shadow models that are trained on the same (or a similar) distribution as the training data. The method assigns membership to input and constructs a new dataset to train the classifier. Subsequently, various MIA attacks were developed considering different threat models and application scenarios. The work in [11] proves that when the adversary has data from a different but similar task, the shadow models are not needed, while a threshold reaching max prediction confidence can provide satisfactory results. The results in [12] find that the training process of ML models is the key to implementing a successful MIA. As the goal is to minimize losses associated with the training samples, members in general tend to have smaller losses than non-member samples. It has been shown that the effectiveness of MIA can be improved by using inputs from query methods [13]. The vulnerability of adversarially trained models to MIA attacks has also been exploited [14].

2.1. Model Quantization

Quantization methods have been shown to be promising in reducing the memory footprint, computational complexity, and energy consumption of neural networks. They focus on converting floating-point numbers into representations with lower bitwidth. For example, quantization can be used to reduce the model size by converting all the parameters' precision from 32 bits to 8 bits or lower for achieving a higher compression rate or acceleration [15]. Extreme quantization is also possible where the model weights can be binary [16] or ternary [17]. In general, quantization methods can be divided into three categories.

Traditional quantization. In these methods, all weights and activations would be quantized. For instance, a non-uniform quantization method uses reduced bitwidth for the majority of data while a small amount of data are handled with high bitwidth [18]. A different approach in the same category utilizes a quantizer that dynamically adapts to the distribution of the parameters [19]. A quantization algorithm is developed by approximating the gradient to the quantizer step size, which can perform comparably to the full bitwidth model [20]. In [21], the proposed quantization function is a linear combination of several sigmoid functions with learnable biases and scales. The method proposed in [16] restricts weights and activations to binary values (−1,1)

, enabling further reduction in memory footprint and efficient hardware implementation. A more stringent quantization method uses three levels (−1,0,1)
to represent weights and activations, striking a balance between binary quantization and full bitwidth.

Mixed-precision quantization. To avoid performance deterioration, some studies suggest using mixed-precision quantization instead of compressing all the layers to the same bidwidth. Mixed-precision quantization typically involves dividing the network into layers or blocks and applying different bitwidths to each part based on its importance and sensitivity to quantization. For example, the quantization bitwidths can be obtained by exploiting second-order (Hessian matrix) information [22]. Differentiable architecture search is also employed by [23][24] to perform mixed-precision quantization.

Dynamic inference quantization. Dynamic inference quantization offers several benefits, including improved flexibility, enhanced adaptability to varying run-time conditions, and potentially better accuracy than quantization with fixed bitwidth. By adjusting the quantization bitwidth on the fly, dynamic inference quantization enables efficient deployment of deep neural network models in resource-constrained environments without sacrificing accuracy. To this end, one approach is to use a bit-controller trained jointly with the given neural network for dynamic inference quantization [25]. Another study [26] proposes dynamically adjusting the quantization interval based on time step information. An algorithm developed by [27] detects sensitive regions and proposes an architecture that employs a flexible variable-speed mixed-precision convolution array.

2.2. Defense against MIA

A defense mechanism against MIA, named MemGuard, was developed [28], which can evade the attacker's membership classification and transform the prediction scores into an adversarial example. MemGuard adds a carefully crafted noise vector to the prediction vector and turns it into an adversarial example of the attack model. Differential privacy [29][30], which can provide a probabilistic guarantee of privacy, has also been shown to be effective in enhancing resistance against MIA [31]. However, differential privacy is costly to implement, and the accuracy reduction makes the method impractical. Distillation for membership privacy (DMP) is a method proposed by [32]. DMP first trains a teacher model and uses it to label data records in the unlabeled reference dataset. The teacher method has no defense mechanism. DMP requires a private training dataset and an unlabeled reference dataset. The purifier framework [33], where the confidence scores of the target model are used as input and are purified by reducing the redundant information in the prediction score, has also been proposed to defend against MIA.

On the other hand, regularization methods designed to reduce overfitting in machine learning models can be employed as defense strategies against MIAs. Adversarial regularization [34] and Mixup + MMD [35] are specific regularization techniques intended to mitigate MIAs. Using regularization, the model generalization is improved and the gap between member and non-member data samples is reduced. **Table 1** summarized prior work based on attack knowledge, MIA attack, and defense mechanism.

Table 1. Prior of each table appears in numerical order. research on defense against MIA.

	Reference	Attack Knowledge	Corresponding Attack	Defense Mechanism
1	[31]	Black-box	Shadow training	Differential privacy
2	[32]	Black-box and White-box	Classifier based and Prediction loss	Distillation
3	[33]	Black-box	Classifier based and Prediction correctness	Prediction purification
4	[34]	Black-box	Shadow training	Regularization
5	[35]	Black-box	Shadow training	Regularization
6	[28]	Black-box	Classifier based	MemGuard

1. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Li, F.F. ImageNet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
2. Ofer, D.; Brandes, N.; Linial, M. The language of proteins: NLP, machine learning & protein sequences. Comput. Struct. Biotechnol. J. 2021, 19, 1750–1758.
3. Wu, J.; Leng, C.; Wang, Y.; Hu, Q.; Cheng, J. Quantized convolutional neural networks for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Los Alamitos, CA, USA, 27–30 June 2016; pp. 4820–4828.
4. Zhou, S.; Ni, Z.; Zhou, X.; Wen, H.; Wu, Y.; Zou, Y. DoReFa-Net: Training Low Bitwidth Convolutional Neural Networks with Low Bitwidth Gradients. arXiv 2016, arXiv:1606.06160.
5. Guo, R.; Sun, P.; Lindgren, E.; Geng, Q.; Simcha, D.; Chern, F.; Kumar, S. Accelerating Large-Scale Inference with Anisotropic Vector Quantization. In Proceedings of the International Conference on Machine Learning, Virtual, 13–18 November 2020; pp. 3887–3896.
6. Giger, M.L. Machine Learning in Medical Imaging. J. Am. Coll. Radiol. 2018, 15, 512–520.
7. Kocić, J.; Jovičić, N.; Drndarević, V. An end-to-end deep neural network for autonomous driving designed for embedded automotive platforms. Sensors 2019, 19, 2064.
8. Prakash, R.M.; Thenmoezhi, N.; Gayathri, M. Face recognition with convolutional neural network and transfer learning. In Proceedings of the International Conference on Smart Systems and Inventive Technology, Tirunelveli, India, 27–29 November 2019; pp. 861–864.
9. Chen, M.X.; Lee, B.N.; Bansal, G.; Cao, Y.; Zhang, S.; Lu, J.; Tsay, J.; Wang, Y.; Dai, A.M.; Chen, Z.; et al. Gmail smart compose: Real-time assisted writing. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 2287–2295.

10. Shokri, R.; Stronati, M.; Song, C.; Shmatikov, V. Membership inference attacks against machine learning models. In Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22–24 May 2017; pp. 3–18.
11. Salem, A.; Zhang, Y.; Humbert, M.; Berrang, P.; Fritz, M.; Backes, M. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. arXiv 2018, arXiv:1806.01246.
12. Liu, Y.; Zhao, Z.; Backes, M.; Zhang, Y. Membership inference attacks by exploiting loss trajectory. In Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, Los Angeles, CA, USA, 7–11 November 2022; pp. 2085–2098.
13. Yeom, S.; Giacomelli, I.; Fredrikson, M.; Jha, S. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. In Proceedings of the 2018 IEEE 31st Computer Security Foundations Symposium (CSF), Oxford, UK, 9–12 July 2018; pp. 268–282.
14. Song, L.; Shokri, R.; Mittal, P. Privacy risks of securing machine learning models against adversarial examples. In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, London, UK, 11–15 November 2019; pp. 241–257.
15. Liu, Z.; Cheng, K.T.; Huang, D.; Xing, E.P.; Shen, Z. Nonuniform-to-uniform quantization: Towards accurate quantization via generalized straight-through estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 4942–4952.
16. Courbariaux, M.; Hubara, I.; Soudry, D.; El-Yaniv, R.; Bengio, Y. Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. arXiv 2016, arXiv:1602.02830.
17. Liu, B.; Li, F.; Wang, X.; Zhang, B.; Yan, J. Ternary weight networks. In Proceedings of the ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
18. Park, E.; Yoo, S.; Vajda, P. Value-aware quantization for training and inference of neural networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 580–595.
19. Baskin, C.; Liss, N.; Schwartz, E.; Zheltonozhskii, E.; Giryes, R.; Bronstein, A.M.; Mendelson, A. Uniq: Uniform noise injection for non-uniform quantization of neural networks. ACM Trans. Comput. Syst. 2021, 37, 1–15.
20. Esser, S.K.; McKinstry, J.L.; Bablani, D.; Appuswamy, R.; Modha, D.S. Learned step size quantization. arXiv 2019, arXiv:1902.08153.
21. Yang, J.; Shen, X.; Xing, J.; Tian, X.; Li, H.; Deng, B.; Huang, J.; Hua, X.S. Quantization networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long

Beach, CA, USA, 16–20 June 2019; pp. 7308–7316.

22. Dong, Z.; Yao, Z.; Arfeen, D.; Gholami, A.; Mahoney, M.W.; Keutzer, K. Hawq-v2: Hessian aware trace-weighted quantization of neural networks. Adv. Neural Inf. Process. Syst. 2020, 33, 18518–18529.
23. Cai, Z.; Vasconcelos, N. Rethinking differentiable search for mixed-precision neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 2349–2358.
24. Wu, B.; Wang, Y.; Zhang, P.; Tian, Y.; Vajda, P.; Keutzer, K. Mixed precision quantization of convnets via differentiable neural architecture search. arXiv 2018, arXiv:1812.00090.
25. Liu, Z.; Wang, Y.; Han, K.; Ma, S.; Gao, W. Instance-aware dynamic neural network quantization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–20 June 2022; pp. 12434–12443.
26. So, J.; Lee, J.; Ahn, D.; Kim, H.; Park, E. Temporal Dynamic Quantization for Diffusion Models. arXiv 2023, arXiv:2306.02316.
27. Song, Z.; Fu, B.; Wu, F.; Jiang, Z.; Jiang, L.; Jing, N.; Liang, X. DRQ: Dynamic Region-based Quantization for Deep Neural Network Acceleration. In Proceedings of the 2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA), Valencia, Spain, 30 May–3 June 2020; pp. 1010–1021.
28. Jia, J.; Salem, A.; Backes, M.; Zhang, Y.; Gong, N.Z. Memguard: Defending against black-box membership inference attacks via adversarial examples. In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, London, UK, 11–15 November 2019; pp. 259–274.
29. Iyengar, R.; Near, J.P.; Song, D.; Thakkar, O.; Thakurta, A.; Wang, L. Towards practical differentially private convex optimization. In Proceedings of the 2019 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 20–22 May 2019; pp. 299–316.
30. Dwork, C.; McSherry, F.; Nissim, K.; Smith, A. Calibrating noise to sensitivity in private data analysis. In Proceedings of the Theory of Cryptography: Third Theory of Cryptography Conference, New York, NY, USA, 4–7 March 2006; pp. 265–284.
31. Chen, Q.; Xiang, C.; Xue, M.; Li, B.; Borisov, N.; Kaarfar, D.; Zhu, H. Differentially private data generative models. arXiv 2018, arXiv:1812.02274.
32. Shejwalkar, V.; Houmansadr, A. Membership privacy for machine learning models through knowledge transfer. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 May 2021; Volume 35, pp. 9549–9557.

33. Yang, Z.; Shao, B.; Xuan, B.; Chang, E.C.; Zhang, F. Defending model inversion and membership inference attacks via prediction purification. arXiv 2020, arXiv:2005.03915.
34. Nasr, M.; Shokri, R.; Houmansadr, A. Machine learning with membership privacy using adversarial regularization. In Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, Toronto, ON, Canada, 15–19 October 2018; pp. 634–646.
35. Li, J.; Li, N.; Ribeiro, B. Membership inference attacks and defenses in classification models. In Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy, Virtual, 26–28 April 2021; pp. 5–16.