TOWARDS MODEL QUANTIZATION ON THE RESILIENCE AGAINST MEMBERSHIP INFERENCE ATTACKS

Charles Kowalski, Azadeh Famili, and Yingjie Lao

Clemson University
Holcombe Department of Electrical and Computer Engineering
Clemson, SC, 29634

2022 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-9620-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICIP46576.2022.9897681

ABSTRACT

As neural networks get deeper and more computationally intensive, model quantization has emerged as a promising compression tool offering lower computational costs with limited performance degradation, enabling deployment on edge devices. Meanwhile, recent studies have shown that neural network models are vulnerable to various security and privacy threats. Among these, membership inference attacks (MIAs) are capable of breaching user privacy by identifying training data from neural network models. This paper investigates the impact of model quantization on the resistance of neural networks against MIA through empirical studies. We demonstrate that quantized models are less likely to leak private information of training data than their full precision counterparts. Our experimental results show that the precision MIA attack on quantized models is 7 to 9 points lower than their counterparts when the recall is the same. To the best of our knowledge, this paper is the first work to study the implication of model quantization on the resistance of neural network models against MIA.

Index Terms— Membership Inference Attack, Model Quantization, Privacy, Neural Network

1. INTRODUCTION

Recent breakthroughs and progress in machine learning have resulted in notable improvements in neural network accuracy for a wide range of tasks [1, 2]. To perform these tasks more effectively, model parameter sizes have expanded drastically. Hence, these networks require excessive computational resources that are not readily available in edge devices, which is one of the main obstacles in neural network deployment. To this end, researchers have proposed quantization methods to compress and speed up training and inference [3, 4, 5] by executing the operations with reduced precision. These techniques have demonstrated comparable performance to full precision networks while being easily accommodated on resource-constrained devices.

Besides edge computing, model quantization along with compression enables broader machine learning applications, including sectors that deal with private data. Notably-sensitive use cases include applications in medical imaging [6], autonomous driving [7], facial recognition [8], and natural language processing [2]. However, as these technologies become increasingly intertwined with daily life, it is imperative that they are continuously evaluated for vulnerabilities and privacy concerns.

Unfortunately, recent studies have demonstrated that machine learning models are quite vulnerable to well-crafted adversarial attacks [9, 10]. For instance, adversarial example attacks can easily use undetectable perturbations to deceive the models and cause misclassifications. Researchers have investigated these attacks and their impact on quantized models [11, 12]. It is reported in [13] that model quantization can help improve the robustness of the model against certain adversarial attacks or even be used as a defensive countermeasure. The effect of model quantization on backdoor and poisoning attacks has also been studied recently [14, 15].

Following the direction of these prior works, this paper evaluates the impact of model quantization on a privacy threat, membership inference attack (MIA), which attempts to identify private user data from a given trained model. To the best of our knowledge, the implication of model quantization on the resistance of neural network models against MIA has not been studied in the literature. We leverage state-of-the-art MIAs used in [16, 17, 18] to compare the effectiveness of the attacks between the quantized and full precision models. Our empirical study shows that the MIA attack accuracy on quantized models never reaches the peak as full precision models do. The same applies to precision and recall. In the case of quantized models, these values fall faster than the full precision models, which means the attack is more likely to result in false positives. From our experimental results, we find that model quantization provides inherent resistance to MIA, along with the reduction in computational complexity. The remainder of this paper is organized as follows. In Section 2, we summarize the background of model quantization and MIA. Then, we describe the details of the methods and implementations in Section 3. Section 4 presents our experimental results. Finally, we conclude our work in Section 5.

2. BACKGROUND

2.1. Model Quantization

The recent achievements in the state-of-the-art performance of neural networks are partially due to a sheer increase in the number of parameters and operations. When dealing with low-power, real-world applications, these costs become enormous obstacles. Thus, model quantization techniques have been developed to compress neural networks, resulting in simpler operations and fewer memory requirements. The existing quantization methods can be broadly categorized into inference quantization methods and quantization-aware training [11]. Inference quantization typically applies the quantization techniques to a pre-trained model [19]. In contrast, the methods in the second category implement quantization during training [3, 4, 5]. For example, the work in [5] uses vector methods such as product and residual quantization to compress models. Although this method produced satisfactory results, it did not fully quantize the models, as this work only focused on the compression of the fully-connected layers. The work in [3] extends the model quantization technique to the convolutional layers and demonstrates the effectiveness on AlexNet. In this paper, we leverage the quantization method proposed in [4], which has yielded excellent performance in a wide range of network architectures.

2.2. Membership Inference Attack

MIA was first introduced by [16], which demonstrated the privacy leakage problem of neural networks. In this work, the authors trained several models called shadow models to imitate the target model's behavior. Then, the adversary could identify whether a given data sample belongs to the target's model training dataset by using the trained shadow models. In the literature, most of the prior works on MIA assume the adversary has access to part or all of the confidence vector of the model. Contrary to the first study, where the shadow models were trained on the same or similar distribution as the target model's dataset, [17] extended the attack strategy using different datasets and a similar task. MIA methods without the need for training shadow models have also been developed [17, 20]. Recently, label-only MIA attacks have also been proposed [21], which can perform the attack without accessing the confidence vector.

3. METHODS

3.1. Quantization

To generate quantized models, we utilize the method developed in [4], DoReFa-Net. While DoReFa-Net was tested only on AlexNet in the original paper, it also yielded excellent performance on other and more recent networks. Its method for weight quantization can be expressed as follows:

$$\textbf{Forward:} \quad r_o = sign(r_i) \cdot S(|r_i|), \tag{1}$$

$$\textbf{Backward:} \quad \frac{\partial f}{\partial r_i} = \frac{\partial f}{\partial r_o} \tag{2}$$

where r_i is a real number input, r_o is a k-bit quantized number, f represents the objective function, and S is a scaling factor. Such transformations are performed through all filters. Additionally, if $k > 1$, DoReFa-Net employs the following transformation:

$$Q^k(r_i) = 2 \cdot Q^k(\frac{\tanh(r_i)}{2 \cdot max(|\tanh(r_i)|)} + \frac{1}{2}) - 1 \tag{3}$$

This equation first limits the values of the weights to $[-1, 1]$ and then quantizes them to the desired k bits within the range $[0, 1]$. Finally, an affine transformation returns the range to the initial $[-1, 1]$. To quantize the activations, DoReFa-Net employs the quantization function, which can be expressed as $a_q = Q^k(a_i)$, where a_i and a_q represent an initial activation and the quantized activation, respectively.

3.2. MIA

The underlying idea behind MIA is that each data point a model was trained on has a distinguishable effect on the model itself: the model should be more confident in classifying training data in general. If we have a pre-trained model h, and f as a decision rule, given data d, the membership prediction can be presented as $f(d; h) \in \{0, 1\}$. We assume that as an adversary, we have access to the trained model h. We can use a query interface on sample data d and then collect the confidence vector $h(x)$. To perform the attack, we query the model. The attack uses an unsupervised binary classification as seen in [17]. Using the maximum value of the extracted vector allows the adversary to compare this value to a certain threshold. If the maximum value for data d is above the threshold, then we consider d a member of the training set $f(d; h) = \{1\}$. The intuition behind the attack is that the maximum value confidence vector of a member data point is much higher than that of a non-member data point.

3.3. Gap Attack

We follow the same setup as in [21] to consider the intuitive assumption made by [22] as a baseline attack, which predicts any misclassified data sample as a non-member of the private training set. This attack is also referred to as a gap attack, since the attack is correlated with the gap between train accuracy and test accuracy. The attack can be expressed as:

$$\frac{1}{2} + \frac{(train_{acc} - test_{acc})}{2} \tag{4}$$

The gap attack can be used as a baseline prediction rather than an actual MIA attack. Since we use quantized models where the training accuracy does not reach 100% training accuracy, there will be misclassified training images that contradict the gap attack intuition.

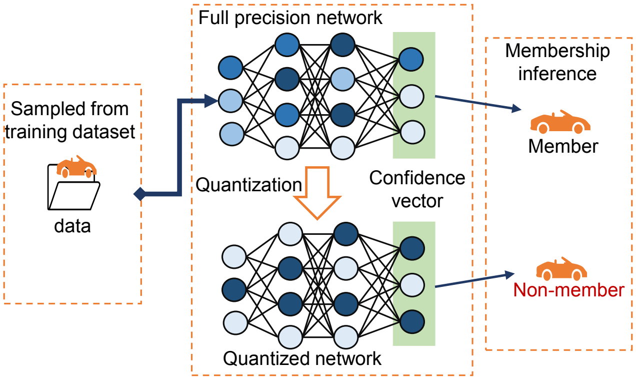

Fig. 1. Overview of the experimental setup.

4. EVALUATIONS

In this section, we perform extensive empirical experiments to answer the following research questions:

1. How does a quantized model respond to MIA in comparison with a full precision model?
2. How does changing a model's activation and bit-width values impact its resilience?
3. Can quantization provide resistance to MIA?

4.1. Threat Model

We consider the same threat model as in prior works [16]. We assume the adversary knows the task and can access partial data samples from a similar distribution to the target model. The adversary can access the softmax layer outputs.

4.2. Experimental Setup

We implement the widely-used ResNet-18 as our full precision baseline. We use the quantization methods described in Section 3.1 to train several quantized versions of ResNet-18. The models are trained and validated with the CIFAR-10 and Fashion MNIST datasets. Since we are not using shadow models, there is no need to split the dataset in the same way as for training. Instead, we split each dataset in half and used one part to train our target model, with the other half left out for testing. In this case, we still see some accuracy degradation. Thus, we use data augmentation to improve the accuracy of our models. The experiments follow a three-step procedure as depicted in Fig. 1. We perform quanizaion to obtain a quantized network. We pass data through full precision and quantized network. Then, we extract the confidence vectors from the networks. Finally, we can assess whether the data sample is a member or non-member based on the confidence vector.

To generate the quantized models, we chose weight bit-width (w) and activation bit-width (a) of 4-bits and 16-bits, respectively. We train three models with these specifications and report their accuracy. For Fashion MNIST, which is a small grey-scale dataset, we use a learning rate of 0.001 for 10 epochs. For the CIFAR-10 dataset, we use a learning rate of 0.01 and multiply the learning rate by 0.1 after 20 epochs. We trained both the full precision and quantized ResNet-18 for 30 epochs. All hyperparameters are kept the same for both training scenarios.

Note that since we only use half of the dataset for training and quantized models, some models don't reach 100% training accuracy. We perform the gap attack on these models. For MIA, we select a range of threshold values and calculate the corresponding $Precision$, $Recall$, and $F1$-Score. We report the highest $Precision$ when $Recall$ is close to 1, and the weighted average of $Precision$ and $Recall$ as $F1$-Score, which can be given by:

$$F1 = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision} \tag{5}$$

4.3. Results

We present the results of the gap attacks in Table 1. For Fashion MNIST, all the models achieved high accuracy. The gap attack accuracy is slightly better than guessing the membership. However, as shown in Fig. 2, the membership inference attack based on confidence vectors is generally worse on quantized networks. When the activation bit is kept at 4-bits during the training phase, they seem resistant to the attack. By changing the threshold value, the attack on the full precision model peaks above 57%, while the heavily quantized models peak at around 54%.

Table 1. Performance comparison of Gap attack on vision based tasks, using full precision ResNet-18, and quantized ResNet-18

Dataset	Test Acc.(%)	Method	Attack Acc.(%)
Fashion MNIST	85.92	full precision	56.00
	89.20	quantized (w=4,a=4)	54.00
	89.08	quantized (w=16,a=4)	54.00
	87.59	quantized (w=16,a=16)	56.03
CIFAR-10	75.68	full precision	62.15
	85.24	quantized (w=4,a=4)	57.32
	84.84	quantized (w=16,a=4)	57.52
	84.93	quantized (w=16,a=16)	57.53

The quantized model with 16-bit weight and activation performs better than the full precision model. This is possibly due to the fact that classifying Fashion MNIST is an easy

Table 2. Performance comparison of confidence vector (CV) based MIA on vision based tasks

Dataset	Test Acc.(%)	Method	$F1$-Score (%)	$Precision$	$Recall$
Fashion MNIST	85.92	CV full precision	66.15	54.00	94.00
	89.20	CV quantized (w=4,a=4)	66.12	51.00	94.00
	89.08	CV quantized (w=16,a=4)	66.09	52.00	94.00
	87.59	CV quantized (w=16,a=16)	68.50	54.00	94.00
CIFAR-10	75.68	CV full precision	77.30	63.00	100.0
	85.24	CV quantized (w=4,a=4)	70.12	54.00	100.0
	84.84	CV quantized (w=16,a=4)	70.96	55.00	100.0
	84.93	CV quantized (w=16,a=16)	71.79	56.00	100.0

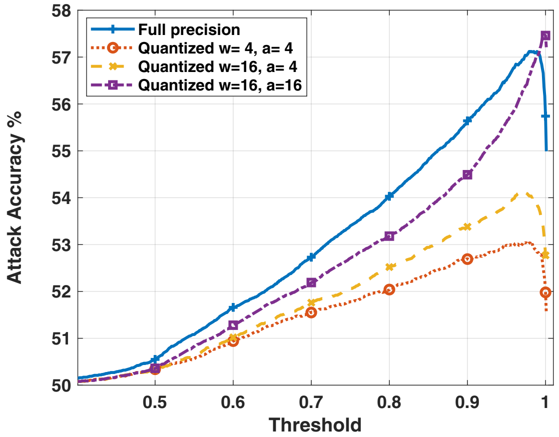

Fig. 2. Accuracy of MIA on quantized ResNet-18 and full precision ResNet-18 on Fashion MNIST datatset.

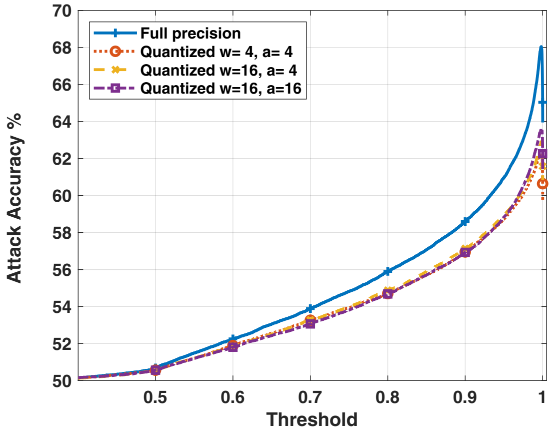

Fig. 3. Accuracy of MIA on quantized ResNet-18 and full precision ResNet-18 on CIFAR-10 dataset.

task. We perform the same experiment on CIFAR-10 as well. We report the results of the gap attack on CIFAR-10 in Fig. 3. As the figure shows, both the full precision and quantized network achieve near 100% training accuracy. The gap attack accuracy indicates that these models are more susceptible to MIA than the previous task. As we can see from Fig. 3, the MIA based on the confidence vector reaches more than 60% for both the full precision and quantized models. The attack accuracy is at its lowest when the activation and weight are restricted to 4-bit during the training phase. Even when the activation and weight are restricted to 16-bit, the attack accuracy is significantly lower than the full precision model. By changing the threshold value, the attack on the full precision model peaks above 68%, while the heavily quantized models peak at around 63%. Since the attack is a binary classification, we also examine other metrics, including true positives and true negatives. When the $Recall$ is high, a high $Precision$ indicates a low number of false positives. Table 2 shows the result for both fashion MNIST and CIFAR-10. We present the highest $F1$-Score with the highest $Recall$ and $Precision$.

Even though quantized networks and full precision networks have similar performance, our experimental results show that quantized networks offer superior protection against MIA. The trend in Fig. 2 where the heavily quantized network provides more protection than the other quantized networks could be a promising direction to enhance the privacy protection of a neural network model.

5. CONCLUSION

This paper studied the implications of model quantization on the privacy leakage of neural network models. To the best of our knowledge, this work is the first to study the relationship between model quantization and its impact on the resistance against MIA. We demonstrated that a quantized model not only reduces the computational complexity from the full precision neural network while maintaining a comparable accuracy, but also provides inherent resistance to MIA.

Acknowledgements

This work is partially supported by the National Science Foundation award 2047384.

6. REFERENCES

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[2] Dan Ofer, Nadav Brandes, and Michal Linial, "The language of proteins: Nlp, machine learning & protein sequences," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 1750–1758, 2021.

[3] Jiaxiang Wu, Cong Leng, Yuhang Wang, Qinghao Hu, and Jian Cheng, "Quantized convolutional neural networks for mobile devices," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4820–4828.

[4] Shuchang Zhou, Zekun Ni, Xinyu Zhou, He Wen, Yuxin Wu, and Yuheng Zou, "Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients," *CoRR*, vol. abs/1606.06160, 2016.

[5] Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar, "Accelerating large-scale inference with anisotropic vector quantization," in *International Conference on Machine Learning*, 2020, pp. 3887–3896.

[6] Maryellen L. Giger, "Machine learning in medical imaging," *Journal of the American College of Radiology*, vol. 15, no. 3, Part B, pp. 512–520, 2018.

[7] Jelena Kocić, Nenad Jovičić, and Vujo Drndarević, "An end-to-end deep neural network for autonomous driving designed for embedded automotive platforms," *Sensors*, vol. 19, no. 9, pp. 2064, 2019.

[8] R Meena Prakash, N Thenmoezhi, and M Gayathri, "Face recognition with convolutional neural network and transfer learning," in *International Conference on Smart Systems and Inventive Technology*, 2019, pp. 861–864.

[9] Joseph Clements and Yingjie Lao, "DeepHardMark: Towards watermarking neural network hardware," in *The Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

[10] Bingyin Zhao and Yingjie Lao, "CLPA: Clean-label poisoning availability attacks using generative adversarial nets," in *The Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

[11] Chang Song, Elias Fallon, and Hai Li, "Improving adversarial robustness in weight-quantized neural networks," *arXiv preprint arXiv:2012.14965*, 2020.

[12] Maungmaung Aprilpyone, Yuma Kinoshita, and Hitoshi Kiya, "Adversarial robustness by one bit double quantization for visual classification," *IEEE Access*, vol. 7, pp. 177932–177943, 2019.

[13] Ji Lin, Chuang Gan, and Song Han, "Defensive quantization: When efficiency meets robustness," *arXiv preprint arXiv:1904.08444*, 2019.

[14] Xudong Pan, Mi Zhang, Yifan Yan, and Min Yang, "Understanding the threats of trojaned quantized neural network in model supply chains," in *Annual Computer Security Applications Conference*, 2021, pp. 634–645.

[15] Hua Ma, Huming Qiu, Yansong Gao, Zhi Zhang, Alsharif Abuadbba, Anmin Fu, Said Al-Sarawi, and Derek Abbott, "Quantization backdoors to deep learning models," *arXiv preprint arXiv:2108.09187*, 2021.

[16] Reza Shokri, Marco Stronati, and Vitaly Shmatikov, "Membership inference attacks against machine learning models," *CoRR*, vol. abs/1610.05820, 2016.

[17] Ahmed Salem, Yang Zhang, Mathias Humbert, Mario Fritz, and Michael Backes, "Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models," *CoRR*, vol. abs/1806.01246, 2018.

[18] Milad Nasr, Reza Shokri, and Amir Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *IEEE Symposium on Security and Privacy*, 2019, pp. 739–753.

[19] Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W Mahoney, and Kurt Keutzer, "A survey of quantization methods for efficient neural network inference," *arXiv preprint arXiv:2103.13630*, 2021.

[20] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro, "Logan: Membership inference attacks against generative models," in *Proceedings on Privacy Enhancing Technologies*, 2019, number 1, pp. 133–152.

[21] Christopher A Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot, "Label-only membership inference attacks," in *International Conference on Machine Learning*, 2021, pp. 1964–1974.

[22] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *IEEE 31st Computer Security Foundations Symposium*, 2018, pp. 268–282.