MEMBERSHIP INFERENCE ATTACKS AND DEFENSES IN NEURAL NETWORK PRUNING

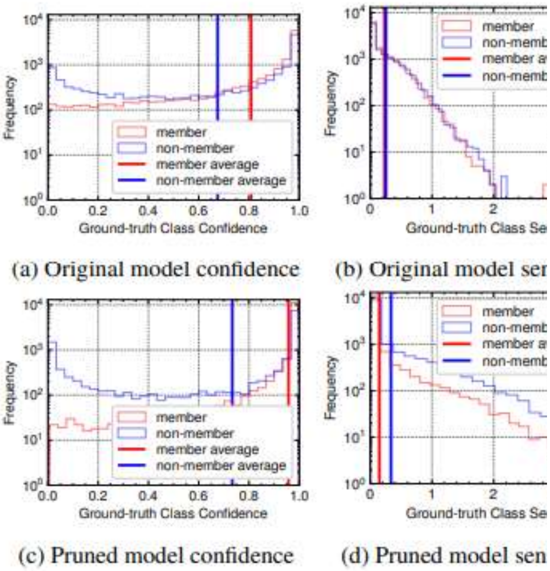
DR. G. KALPANA$_1$, A. RENUSRI$_2$, AYESHA BEGUM$_3$, D. NAVYA TEJA$_4$

1ASSOSCIATE PROFESSOR, DEPARTMENT OF CSE, MALLA REDDY ENGINEERING COLLEGE FOR WOMEN, HYDERABAD.

2,3&4UG SCHOLAR, DEPARTMENT OF CSE, MALLA REDDY ENGINEERING COLLEGE FOR WOMEN, HYDERABAD

ABSTRACT Neural network pruning has been an essential technique to reduce the computation and memory requirements for using deep neural networks for resource-constrained devices. Most existing research focuses primarily on balancing the sparsity and accuracy of a pruned neural network by strategically removing insignificant parameters and retraining the pruned model. Such efforts on reusing training samples pose serious privacy risks due to increased memorization, which, however, has not been investigated yet. In this paper, we conduct the first analysis of privacy risks in neural network pruning. Specifically, we investigate the impacts of neural network pruning on training data privacy, i.e., membership inference attacks. We first explore the impact of neural network pruning on prediction divergence, where the pruning process disproportionately affects the pruned model's behavior for members and non-members. Meanwhile, the influence of divergence even varies among different classes in a fine-grained manner. Enlightened by such divergence, we proposed a self-attention membership inference attack against the pruned neural networks. Extensive experiments are conducted to rigorously evaluate the privacy impacts of different pruning approaches, sparsity levels, and adversary knowledge. The proposed attack shows the higher attack performance on the pruned models when compared with eight existing membership inference attacks. In addition, we propose a new defense mechanism to protect the pruning process by mitigating the prediction divergence based on KL-divergence distance, whose effectiveness has been experimentally demonstrated to effectively mitigate the privacy risks while maintaining the sparsity and accuracy of the pruned models.

INTRODUCTION Much of the progress in artificial intelligence over the past decade has been the result of deep neural networks (DNNs). The powerful DNNs with a large number of parameters consume considerable storage and memory bandwidth, which makes it challenging to deploy the state-of-the-art neural networks on resource-constrained devices. To address this issue, neural network pruning as one of the most popular compression technologies has attracted great attention [1, 2]. By removing insignificant parameters from a DNN, recent research has shown that neural network pruning can substantially reduce the size of a DNN and speedup the inference process without largely compromising prediction accuracy [2–5]. In general, neural network pruning includes three main stages: 1) train an original DNN; 2) remove the insignificant parameters; 3) fine-tune the remaining parameters with the training dataset. Most existing research on neural network pruning has focused on improving the trade-off between accuracy and sparsity by strategically designing the last two stages [2–5]. However, such efforts on reusing training samples pose serious privacy risks of the pruned neural networks due to the potentially increased memorization of training samples. The privacy risks of DNNs have already been pointed out, where a DNN is prone to memorizing sensitive information of the training dataset [6–9]. Taking the membership inference attack (MIA) as an example, an adversary can infer whether a given data sample was used to train a DNN, seriously threatening individual privacy. For instance, an adversary can infer an individual was a confirmed case, if it is known that the individual's record was used to train an infectious disease model. The MIA was first proposed against black-box models in [10], where the adversary only has access to the data sample and predictions of the target model. Later on, more attention has been attracted against various DNN models, such as generative models [7, 8], graph models [11], machine translation [12], text generation [13], genomic analysis [14], and transfer learning [15]. Although extensive analysis has been conducted, none of the existing efforts have been put into analyzing MIAs against pruned neural networks. In view of this, the paper focuses on one fundamental question: comparing with original deep neural networks, are the pruned networks more vulnerable to membership inference attacks? Specifically, most MIAs infer a sample's membership based on the different behaviors of a target model between



(a) Original model confidence (b) Original model ser

(c) Pruned model confidence (d) Pruned model sen

Figure 1: Histograms of the prediction confidences and the prediction sensitivity of the ground-truth label. We remove 70% of the parameters in the original DenseNet121 model using l1 unstructured pruning on the CIFAR10 dataset. The figures show the frequency of prediction confidence (a) and (c) and prediction sensitivity (b) and (d) belonging to the groundtruth class on the training and test data. The vertical lines indicate the average values of training data, i.e., members (black), and test data, i.e., non-members (red), respectively. In both prediction confidence and sensitivity measurements, neural network pruning makes the distances between the two vertical lines in the pruned model larger than that in the original model, which indicates a larger confidence gap and sensitivity gap between members and non-members due to pruning. members (i.e., training samples) and non-members (i.e., test samples), such as the different prediction confidences [9, 10]. Since most neural network pruning approaches rely on reusing the training dataset to fine-tune the parameters after pruning the insignificant parameters, the additional training at the pruned neural network inevitably increases its memorization of the training samples. Moreover, the pruned neural network enforces a small number of parameters to achieve similar prediction capabilities, which also increases the memorization of training data and makes the pruned model more sensitive to the training data. Hence, such increased memorization can intuitively lead to a larger divergence of the prediction confidences and sensitivities between members and non-members. Figure 1 illustrates the prediction confidence and the prediction sensitivity1 of members and non-members in the original DNN and the pruned network, respectively. The larger divergence of the confidences and the sensitivities in the pruned model at (c) and (d) confirms our intuition: neural network pruning can aggravate the privacy issues of the original deep neural network. Therefore, in the following paper, we conduct a comprehensive analysis to reveal the impacts of neural network pruning on

training data privacy, i.e., MIAs. Specifically, we first explore the impact of neural network pruning on prediction divergence: the pruning process disproportionately affects the pruned model's behavior for members and nonmembers. Enlightened by this insight, a new MIA is proposed against the pruned neural networks. In addition, with the proposed new attack, we propose a new defense mechanism to protect the fine-tuning process by mitigating the prediction divergence based on KL-divergence distance. Extensive experiments are conducted to rigorously evaluate our proposals. To the best of our knowledge, this is the first study to investigate the privacy risks of neural network pruning. Our main contributions are summarized below: • We investigate the privacy risk of neural network pruning and propose a new MIA: self-attention membership inference attack (SAMIA). By exploring the impacts of neural network pruning on prediction divergence, the proposed attack results in high attack accuracy of revealing the membership status from the pruned models. In particular, SAMIA has advantages in identifying the pruned models' prediction divergence by using finergrained prediction metrics. We recommend SAMIA as a competitive baseline attack model for future privacy risk study of neural network pruning. • To rigorously evaluate the privacy impacts of different pruning approaches, sparsity levels, and adversary knowledge, we conduct extensive experiments on seven commonly used datasets, four neural network architectures, four pruning approaches, five sparsity levels, and 255 pruned models in total. Experimental results demonstrate the effectiveness of the proposed attacks against pruned neural networks, which further indicates that neural network pruning can aggravate the privacy issues of the original DNN. The adversary can successfully reveal the membership status, even without the knowledge of the pruning approach used in the target model. Furthermore, we evaluate the privacy impacts of different pruning approaches and various sparsity levels. • To defend the pruned models against MIAs, we propose a new defense mechanism: pair-based posterior balancing (PPB). PPB protects the fine-tuning process of neural network pruning by narrowing down the divergences of posterior predictions and reducing the prediction sensitivities based on their KL-divergence distances. Experimental results demonstrate the effectiveness of the PPB mechanism, which significantly mitigates the privacy risks while maintaining the sparsity and accuracy of the pruned model.

Besides, compared with the state-of-theart defenses, PPB achieves a better trade-off between prediction performance and privacy in most cases.

BACKGROUND AND RELATED WORK

Neural Network Pruning The state-of-the-art neural networks are usually deep and resource hungry, requiring large amounts of computation and memory, which becomes a particular challenge on resourceconstrained end devices. As one of the most popular network compression approaches, neural network pruning has attracted great attention in recent years [2–5]. In general, most network pruning studies follow the pruning workflow: "train-prunefinetuning." For example, Han et al. [2] proposed to remove the individual parameters with the lowest magnitude. Randomly removing individual parameters reduces the model size, but may not be efficient to facilitate hardware optimization and accelerate the neural network computation. Therefore, many methods were proposed to remove parameters in an organized way by removing a group of parameters (i.e., structured pruning). For example, Li et al. [3] removed the entire filters with the lowest magnitude in the neural network, which leads to significant speedup compared with the unstructured pruning. Liu et al. [4] removed the entire channels according to the corresponding scaling factors in the followed batch normalization layers. In this paper, we investigate the privacy risks of both unstructured and structured pruning approaches. More recently, new pruning approaches have been proposed, which prune parameters by searching the optimal neural architecture [16, 17] or fine-tune the pruned model by rewinding the parameters to the previous states [18, 19]. The privacy risks discussed in this paper might exist in these new pruning approaches. We will investigate their privacy risks in our future work. On the other hand, recent efforts have been put into neural network pruning from other important perspectives. Paganini [20] investigated the unfairness and systematic biases in the pruned models. Hooker et al. [21] demonstrated the biased performance on different groups and classes after pruning. Given the potential of pervasively implementing neural network pruning, this work targets another critical and urgent aspect regarding neural network pruning, i.e., training data privacy.

Membership Inference Attacks (MIAs) Membership inference attacks have raised serious privacy threats by determining if a record was in the training dataset of a neural network model via querying that model. Given a target neural network model f : R n → R, the process of MIA can be formally defined as: A : x, f → {0,1}, (1) where A denotes the attack model, which is a binary classifier. If the data sample x is used to train the target model f , the attack model A outputs 1 (i.e., member), and 0 otherwise (i.e., non-member). Due to the practical consideration, most MIAs focused on the black-box setting, where an adversary only has access to the target model's outputs. By leveraging the target model's prediction confidences, Shokri et al., [22] proposed a blackbox MIA. They constructed several shadow models to mimic the behavior of a target model. The well-established shadow models will then be used to generate data to train a neural network-based binary classifier to determine the membership of a record against the target model, i.e., whether a record belongs to the target model's training dataset or not. Salem et al., [23] further boosted this attack successfully by only using a single shadow model. To further improve the attack accuracy, Nasr et al., [24] included more features, such as the class labels of data samples, to train the binary classifier. In addition to the aforementioned neural network-based binary classifier, Leino et al., [25], Yeom et al., [26], and Song et al., [27,28] proposed the metric-based binary classifier, where the membership of a record is directly determined by a predefined threshold based on the metrics, such as the prediction confidences, entropy, or modified entropy of the record. Song and Mittal showed that by setting a class-dependent threshold, the metric-based classifier could achieve comparable or even better accurate inference performance compared with the neural network-based classifier [28]. Despite the extensive research on MIAs, none of them is designed towards pruned models. Therefore, we propose SAMIA to investigate the privacy risks of pruned models.

Defenses against MIAs Recent efforts have been made to defend against MIAs. As one of the most popular privacy-preserving techniques, differential privacy (DP) provides provable defense against MIAs by adding noise to the gradient or parameter during model training [29–31]. However, DP usually requires a large magnitude of noises to achieve a meaningful privacy guarantee, which seriously degrades the performance of the protected models [32]. On the other hand, regularization [10], dropout, and model stacking [23] have been used in model training to reduce the privacy risks caused by overfitting. Although these approaches reduced the vulnerability by bridging the generalization gap between member and non-member data samples, in many cases, the privacy risks after applying these approaches are still high. Recent adversarial learning techniques [33, 34] have been introduced in defending against MIAs by adding noises to the prediction confidences for misleading the adversary [24, 35]. In a recent analysis of the defense mechanisms, Song and Mittal showed that the early stopping mechanism achieved comparable performance with most defenses [28]. In this paper, we provide a comprehensive analysis of defenses in neural network pruning, including our proposed PPB defense along with the existing defense mechanisms.

ATTACK EVALUATION This section conducts comprehensive experiments2 to thoroughly investigate the privacy risks of the proposed MIAs against neural network pruning. In the following, we first introduce the experimental setup, and then evaluate the privacy risks of the pruned models by comparing them with those of original models. Next, we investigate the impact of the confidence gap, sensitivity gap, and generalization gap, respectively. Finally, we evaluate the privacy risks without the knowledge of pruning approaches and sparsity levels.

Evaluation Setup In the evaluation, we consider the most widely used datasets, neural network architectures, and optimization approaches following recent research of MIAs [10, 23, 28, 45].

Datasets We consider seven popular datasets in the experiments: CIFAR10, CIFAR100, CHMNIST, SVHN, Texas, Location, and Purchase.

- CIFAR10 and CIFAR100 [46]. These are two benchmark datasets for image classification. CIFAR10 dataset contains 60,000 32×32 color images in 10 classes, with 6,000 images per class. CIFAR100 dataset contains 60,000 color images in 100 classes, with 600 images per class.

- CHMNIST [47]. This dataset consists of 5,000 histological images of human colorectal cancer containing 10 classes of tissues. We resize all images to 32×32, the same dimension as CIFAR10 and CIFAR100.

- SVHN [48]. This dataset consists of 99,289 32×32 color images from house numbers in the Google Street View dataset, containing 10 classes from 0 to 9.

- Location [49,50]. This dataset contains location “check-in” records of mobile users in the Foursquare social network, restricted to the Bangkok area. The dataset is used to predict users’ geosocial type based on the geographical history record features: whether the user visited a certain region or location type. We use the preprocessed purchase dataset provided by Shokri et al. [10], which contains 5,010 data samples, 446 binary features, and 30 classes.

• Texas [51]. This dataset is presented in the Hospital Discharge Data Public Use Data File provided by the Texas Department of State Health Services. The dataset is used to predict the types of patient’s main procedure based on a wide range of features, such as external causes of injury, diagnosis of the patient, procedures the patient underwent, and other generic information. We use the preprocessed purchase dataset provided by [10], which contains 67,330 data samples, 6,169 binary features, and 100 classes.

• Purchase [52]. This dataset is presented in Acquire Valued Shoppers Challenge to predict which shoppers will become repeat buyers based on the purchase history. We use the preprocessed purchase dataset provided by Shokri et al. [10], which contains 197,324 data samples, 600 binary features, and 100 classes.

Each above dataset is first randomly and equally split into two parts: one for target model, one for shadow model. In each part, we split the data into three datasets: training (45%), validation (10%), and test (45%). We use the validation dataset to determine if the model needs to stop training or fine-tuning for early stopping. Therefore, the membership inference via random guessing results in 50% attack accuracy

CONCLUSION

This paper conducted the first analysis of privacy risks in neural network pruning. We first explored the impacts of neural network pruning on prediction divergence, based on which, a new membership inference attack, i.e., self-attention membership inference attack (SAMIA), is proposed against the pruned neural network models. Through comprehensive and rigorous evaluation, we demonstrated the substantially increased privacy risks of the pruned models. We found that the privacy risks of the pruned models are tightly related to the confidence gap, sensitivity gap, and generalization gap due to pruning. Besides, even without knowing the pruning approach, the membership inference attacks can still achieve high attack accuracy against the pruned model. Especially, the proposed SAMIA showed superiority in identifying the pruned models’ prediction divergence by using finer-grained prediction metrics, which is recommended as a competitive baseline attack model for future privacy risk study of neural network pruning. In addition, to defend the attacks, we proposed a pair-based posterior balancing named as PPB by reducing the prediction divergence of fine-tuning process during neural network pruning. We experimentally demonstrated that PPB could reduce the attack accuracy to around 50% (random guessing accuracy) without considering adaptive attacks and achieve the best protection compared with the three existing defenses. Besides, PPB showed competitive performance even when defending adaptive attacks. The proposed SAMIA attack will be further explored under more challenging MIA settings, such as the label-only MIA without available confidences, where the existing label-only MIA attacks using data augmentation [60] and black-box adversary [61] can be potentially integrated for more powerful attack capability. We hope our work convinces the community about the importance of exploring innovative neural network pruning approaches by taking privacy-preserving into consideration.

REFERENCES

[1] Michael Mozer and Paul Smolensky. Skeletonization: A technique for trimming the fat from a network via relevance assessment. In Advances in Neural Information Processing Systems (NIPS), 1988.

[2] Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding. In International Conference on Learning Representations (ICLR), 2016.

[3] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. In International Conference on Learning Representations (ICLR) (Poster), 2017.

[4] Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shoumeng Yan, and Changshui Zhang. Learning efficient convolutional networks through network slimming. In IEEE International Conference on Computer Vision (ICCV), 2017.

[5] Davis W. Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle, and John V. Guttag. What is the state of neural network pruning? In MLSys, 2020.

[6] 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.

[7] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. LOGAN: membership inference attacks

against generative models. Proc. Priv. Enhancing Technol., 2019. [8] Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. Gan-leaks: A taxonomy of membership inference attacks against generative models. In ACM SIGSAC Conference on Computer and Communications Security (CCS), 2020.

[9] Jiacheng Li, Ninghui Li, and Bruno Ribeiro. Membership inference attacks and defenses in classification models. In Conference on Data and Application Security and Privacy (CODASPY), 2021.

[10] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In IEEE Symposium on Security and Privacy, 2017.

[11] Iyiola E. Olatunji, Wolfgang Nejdl, and Megha Khosla. Membership inference attack on graph neural networks. arXiv preprint arXiv:2101.06570, 2021.

[12] Sorami Hisamoto, Matt Post, and Kevin Duh. Membership inference attacks on sequence-to-sequence models: Is my data in your machine translation system? Trans. Assoc. Comput. Linguistics, 2020.

[13] Congzheng Song and Vitaly Shmatikov. Auditing data provenance in text-generation models. In ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2019.

[14] Junjie Chen, Wendy Hui Wang, and Xinghua Shi. Differential privacy protection against membership inference attack on machine learning for genomic data. In PSB, 2021.

[15] Yang Zou, Zhikun Zhang, Michael Backes, and Yang Zhang. Privacy analysis of deep learning in the wild: Membership inference attacks against transfer learning. arXiv preprint arXiv:2009.04872, 2020.