Against Membership Inference Attack: Pruning is All You Need

Yijue Wang	extsuperscript{1}, Chenghong Wang	extsuperscript{2}, Zeng Wang	extsuperscript{1}, Shanglin Zhou	extsuperscript{1}, Hang Liu	extsuperscript{3}, Jinbo Bi	extsuperscript{1}, Caiwen Ding	extsuperscript{1}, Sanguthevar Rajasekaran	extsuperscript{1*}

	extsuperscript{1}University of Connecticut

	extsuperscript{2}Duke University

	extsuperscript{3}Stevens Institute of Technology

\{yijue.wang, zeng.wang, shanglin.zhou, jinbo.bi, caiwen.ding, sanguthevar.rajasekaran\}@uconn.edu, cw374d@duke.edu, hang.liu@stevens.edu

Abstract

The large model size, high computational operations, and vulnerability against membership inference attack (MIA) have impeded deep learning or deep neural networks (DNNs) popularity, especially on mobile devices. To address the challenge, we envision that the weight pruning technique will help DNNs against MIA while reducing model storage and computational operation. In this work, we propose a pruning algorithm, and we show that the proposed algorithm can find a subnetwork that can prevent privacy leakage from MIA and achieves competitive accuracy with the original DNNs. We also verify our theoretical insights with experiments. Our experimental results illustrate that the attack accuracy using model compression is up to 13.6% and 10% lower than that of the baseline and Min-Max game, accordingly.

1 Introduction

Advances in Machine Learning (ML) have enabled high accuracy in classifications, recommendations, and natural language processing, etc [He et al., 2016; Vaswani et al., 2017; Pan et al., 2020]. The success of modern deep neural networks (DNNs) is mainly dependent on the availability of advanced computing power and a large number of data. Machine-Learning-As-A-Service (MLaaS) [Ribeiro et al., 2015] providers such as Amazon [Kurniawan, 2018], Microsoft [Gollob, 2015], and Google [Ravulavaru, 2018] have taken advantage of the aforementioned two availabilities. By providing black-box interfaces, MLaaS allows individuals or groups to upload data easily, leverage powerful large-scale DNNs, and deploy analytic services via pay-as-you-go or subscriptions using personal computers or mobile devices [Truex et al., 2019].

However, there are two main challenges. (i) MLaaS raises privacy concerns on sensitive data such as patient treatment records. Even though the DNN model structures are in black-box, MLaaS can leak sensitive information about training data used to build back-end models. For instance, membership inference attack (MIA) [Shokri et al., 2017] is one of the critical inference attacks in exploiting the aforementioned vulnerability. By using MIA, the adversary monitors the distinctive behavior of back-end models by repeating sophisticated designed inference requests to further exploit information about the training data. (ii) DNN models are evolving fast in order to satisfy the diverse characteristics of broad applications. As the layers of DNNs get deeper and the model size of DNNs gets larger, the large computational operations and model size introduce substantial data movements, limiting their ability to provide a user-friendly experience on mobile devices [Krizhevsky et al., 2012; Hinton et al., 2012].

There are several mechanisms have been developed to address the MIA challenges. Differential privacy (DP), a major privacy-preserving mechanism against general Inference attack, which is based on adding noises into gradients or objective function of the training model, has been applied in different machine learning models [Abadi et al., 2016; Zhang et al., 2019; Rahman et al., 2018]. Although the robustness of DP has been proven, the utility cost (e.g., creating indistinguishable non-membership datasets, calculating the bound for the function sensitivity) of DP is hard to be limited as acceptable since it imposes a significant accuracy loss for protecting complicated models as well as on high dimensional data when noise is considerable.

Another defense mechanism is game theory, e.g., Min-Max game [Nasr et al., 2018], which guarantees the information privacy. The maximum gain of inference model is considered as a new regularization called \textit{adversarial regularization} and will be minimized with the training model loss. Unfortunately, the Min-Max game introduces extra computational operations in addition to the classifier training process.

Finally, yet importantly, neither DP nor Min-Max game addresses the second challenge, i.e., large computational operations and model size in DNNs.

In this paper, to simultaneously address the two challenges, we develop our pruning algorithm that is optimized for the dual objectives of privacy and efficiency by finding a subnetwork from a sufficiently over-parameterized random network. We present the main contributions of our work as follows:

- To the best of our knowledge, this work is the first attempt...
to simultaneously address the challenges of large model size, high computational cost, and vulnerability against MIA on DNNs. We jointly formulate weight pruning and MIA as MIA-Pruning and provide an analytic solution strategy for the problem.

• We show that our pruning algorithm can find a subnetwork that can prevent the privacy leakage from MIA and achieves competitive accuracy with the original DNNs.

• We show that our pruning algorithm performs better than baseline (without defense and pruning) and Min-Max game, i.e., further reduces attack accuracy. We also investigate the combination of weight pruning and Min-Max game and show that the combination will further enhance DNN model privacy.

Experimental results show that our MIA-Pruning can help against MIA while simultaneously achieving model storage and computational complexity reduction within a very small accuracy loss. Our proposed method significantly outperforms DP on MIA. Since weight pruning reduces the number of parameters, the proposed MIA-Pruning enables faster DNN computation than prior works.

2 Related Work

2.1 Weight Pruning

State-of-the-art DNNs contain multiple cascaded layers and at least millions of parameters (i.e., weights) for the entire model [He et al., 2016; Vaswani et al., 2017]. Prior works have focused on developing DNN weight pruning algorithms such as weight pruning [Zhang et al., 2018; Frankle and Carbin, 2018; Zhou et al., 2019; Ramanujan et al., 2020] (i.e., removing weights with specific dimensions or with any desired weight matrix shapes) utilizing different regularization techniques to explore sparsity. Blalock [Blalock et al., 2020] introduced ShrinkBench for standardized evaluations of pruning methods.

2.2 Defense Mechanism against MIA

One defense direction is using game theory to protect privacy [Nasr et al., 2018; Alvim et al., 2017; Shokri, 2015; Shokri et al., 2012]. Most of the game theory-based mechanisms minimize the privacy loss against the most potent attacker by converting the utility function into the Min-Max optimization problem.

[Nasr et al., 2018] proposed a Min-Max game mechanism and formulated the gain of MIA as a new regularization, which is maximized while the classifier’s loss is minimized. We use it as a comparison with our experimental results.

DP is another major defense mechanism against MIA. There are multiple DP-based defense mechanisms [Chaudhuri et al., 2011; Abadi et al., 2016; Iyengar et al., 2019], by adding noises into gradients or the objective function of the training model. However, the existing mechanisms would impose a significant accuracy loss for protecting complicated models as well as on high dimensional data when the noise parameter ε is large.

There are some other defense directions. For example, the model stacking [Salem et al., 2018] mechanism made a combination of multiple classifier results to prevent the attacker from inferring a single target classifier. MemGuard mechanism [Jia et al., 2019] randomly added noise on the target classifier prediction.

The existing defenses have at least one of the limitations: 1) they have typical extra computations, such as extra weight storage and noise calculations. That means these mechanisms introduce extra computational operations in addition to the training approaches. 2) they achieve privacy protection with significant utility loss.

3 MIA-Pruning: Problem Statement

In this work, we investigate the following question: Will an effective DNN weight pruning technique help against MIA while simultaneously achieving model storage and computational complexity reduction?

3.1 Problem Formulation

MIA

For the target machine learning model, we consider the classification model in this work. Let f denotes the target classification model, x denotes a data point, and f(x) denotes the output of f on data x. f(x) is a one-hot vector of probabilities of x belonging to k classes. We consider the MIA problems in a black-box condition, which means the adversary can not access the classification model’s parameters but can only observe the input and output of the classification model. We assume that the adversary has access to some data records from the training set and the predictions from the black-box DNN target model. Based on the difference between the model’s prediction on the training dataset and the non-training dataset, the adversary can determine whether a data record belongs to the training set or not. Figure 1 shows an illustrative diagram of using weight pruning against MIA in DNNs. We use fA to denote the adversarial inference model fA : x × y × f(x) → [0, 1]. fA takes the feature of the data x, the label of the data y, and the prediction of classification model f(x) as inputs. And fA outputs the probability of data (x, y) belonging to the training set D or the non-training set D'. The probability distributions of samples in D and D' are PD and PD', respectively. The gain function of the inference
model $f_A$ given the classification model $f$ can be written as:

$$G_f(f_A) = \mathbb{E}_{(x,y) \sim P_D}[\log(f_A(x,y,f(x)))]
+ \mathbb{E}_{(x,y) \sim P_{D'}}[\log(1 - f_A(x,y,f(x)))]$$

(1)

The first expectation computes the inference model’s accuracy in predicting training data (members), and the second expectation computes the accuracy of the inference model on predicting non-training data (non-members). The underline probability $P_D$ and $P_{D'}$ is normally not known. The empirical gain can be calculated by simply sampling data from the training set and validation set.

**MIA-Pruning against MIA**

The objective is to build a defense system against MIA by pruning the model weights so that the model predictions for the training dataset and non-training dataset are distinguishable. In this case, it becomes more difficult for the adversary to determine where an observed data record belongs to. Finally, the risk of membership privacy loss is reduced. Ideally, the adversary can only make a determination by random guess. At the same time, the classification accuracy of the model will not be or slightly be affected. In other words, the utility cost of defense (e.g., classification accuracy loss) is negligible. Formally, the problem can be formulated as:

$$\arg\min_{\{W_i\}, \{b_i\}} \mathcal{L}(f(\{W_i\}, \{b_i\}; x), y)
\text{s.t. } W_i \in \{W_i | \text{card}(W_i) \leq n_i \}
\{n_i \} = \arg\min _{n_i} \max f_A(W_i, \{b_i\})(f_A)$$

(2)

where $\{W_i\}$ and $\{b_i\}$ are the weights and biases of each layer, and $\mathcal{L}(f(\{W_i\}, \{b_i\}; x), y)$ is the loss of the classification model. $\text{card}(W_i)$ is the cardinality of weights in each layer, which returns the number of non-zero weights. $n_i$ is the desired number of non-zero weights for each layer, which regularizes the strength of compression.

**MIA-Pruning & Min-Max Game against MIA**

For the further step, we consider the combination of weight pruning and Min-Max game [Nasr et al., 2018] to strengthen the defense efficiency against MIA, and the optimization objective would become:

$$\arg\min_{\{W_i\}, \{b_i\}} \mathcal{L}(f(\{W_i\}, \{b_i\}; x), y)
+ \gamma \max f_A(W_i, \{b_i\})(f_A)
\text{s.t. } W_i \in \{W_i | \text{card}(W_i) \leq n_i \}
\{n_i \} = \arg\min _{n_i} \max f_A(W_i, \{b_i\})(f_A)$$

(3)

where $\gamma$ is the weight of the maximum gain of the inference model. The difference between Equation 2 and 3 is that the loss in 3 when updating the classification model also includes the maximum possible inference gain given the current classification model. The problem of 2 and 3 can be written as Equation 14 in Appendix Section 2.

**4 MIA-Pruning: Methodology**

**4.1 Solution Strategy**

To find the optimal compression ratio or $n_i$, we set different $n_i$ in a range and choose $n_i$ that can achieve the minimum $G_f(f_A)$. For the Min-Max game, the minimization and maximization problems need to be solved jointly to reach an equilibrium point. As shown in algorithm 1, for a fixed classification model $f$, we use a subset of training data and non-training data to train the inference model $f_A$ to find the best attack to $f$. Then for the current $f_A$, we minimize the loss in Equation 2 and 3 to find the most defensive model $f$ against $f_A$. Note that for MIA-Pruning, the for-loop to update $f_A$ is not needed. For an arbitrary $f_A$, it is easy to optimize it empirically by Stochastic Gradient Descent (SGD). However, it is difficult to solve the optimization problem of $f$. We modified the Alternating Direction Method of Multipliers(ADMM) [Zhang et al., 2018] to solve it, of which the details can be found in Appendix Section 3.

**4.2 Theoretical Analysis**

**Pruning Convergence Analysis**

We suppose the target network $f(x)$ as:

$$f(x) = W_n^f \sigma(W_{n-1}^f \ldots (\sigma(W_1^f x)))$$

(4)

and we define the original network $g(x)$ as:

$$g(x) = W_{2n}^g \sigma(W_{2n-1}^g \ldots \sigma(W_1^g x))$$

(5)

where $W_i^f, W_j^g$ is the randomized weight matrix at $i$-th layer of $f$ and $j$-th layer of $g(x)$. And $\sigma(\cdot)$ is the activation function.

A pruned network $\hat{g}(x)$ can be presented as:

$$\hat{g}(x) = (P_{2n} \odot W_{2n}^g) \sigma(P_{2n-1} \odot W_{2n-1}^g) \ldots \sigma(P_1 \odot W_1^g x)$$

(6)

where $P_i$ is the pruning matrix in $i$-th layer.

**Theorem 1.** For every network $f$ defined in Eq. 4 with depth $d$ and $\forall i \in \{1,2,\ldots,n\}$. Consider $g$ defined in Eq.5 is a randomly initialized neural network with $2n$ layers, and width $\text{poly}(d,n,m,1/\epsilon, \log 1/\delta)$, where $d$ is input size, $n$ is number of layers in $f$, $m$ is the maximum number of neurons in a layer. The weight initialization distribution belongs to uniform distribution in range [-1, 1]. Then with probability at least $1 - \delta$ there is a weight-pruned subnetwork $\hat{g}$ of $g$ such that:

$$\sup_{x \in \mathcal{X}, \|W\| \leq 1} \|f(x) - \hat{g}(x)\| \leq \epsilon$$

(7)

The full proof of Theorem 1 is in Appendix Section 1. Using Theorem 1, we know that for every bounded distribution and every target network with bounded weights, there is a subnetwork with the competitive accuracy of the original sufficiently over-parameterized neural networks.

**Inference Model’s Gain Function Analysis**

According to [Nasr et al., 2018], we rewrite the gain function of the inference model in the form of probability distribution:

$$G_f(f_A) = \int_{x,y} [P_{D}(x,y) p_f(f(x)) \log(f_A(x,y,f(x)))
+ P_{D'}(x,y) p'_f(f(x)) \log(1 - f_A(x,y,f(x)))]dx dy$$

(8)
Algorithm 1 The Process of MIA-Pruning

1: for \( n_i \) in \( \{ n_{\text{min}}, n_{\text{max}} \} \) do
2: \[ \text{for epoch in epochs do} \]
3: \[ \text{for } t \text{ in iterations do} \]
4: \[ \text{Get a random mini-batch } S \subseteq D, \]
5: \[ \text{Get a random mini-batch } S' \subseteq D', \]
6: \[ \text{Update } f_A \text{ to minimize } -G_f(f_A) \text{ using SDG.} \]
7: \[ \text{end for} \]
8: \[ \text{Get a random mini-batch } S'' \subseteq D, S'' \neq S. \]
9: \[ \text{Update } \{ W_i \}, \{ b_i \} \text{ to minimize Loss in Equation 2} \]
10: \[ \text{Prune weight by update } \{ W_i \} \text{ to } \{ W_i' \} \]
11: \[ \text{end for} \]
12: \[ \text{OUTPUT } \{ W_i \}, \{ b_i \}, G_f(f_A). \]
13: \[ \text{end for} \]
14: \[ \text{OUTPUT } \{ W_i \}, \{ b_i \}, G_f(f_A) = \min(G(f_A)). \]

where \( D \) is the training set and \( D' \) is the non-training set. \( p_f \) and \( p_f' \) are the probability distribution of the classification model \( f \)'s output for training data and non-training data.

For a given classification model \( f \) and data sampled from a known probability distribution, the optimal determination solution for the inference model \( f_A \) is [Goodfellow et al., 2014; Nasr et al., 2018]:

\[
f_A^*(x, y, f(x)) = \frac{p_f(f(x))}{p_f(f(x)) + p_f'(f(x'))}
\]

Therefore, by substituting \( f_A^* \) in the Equation 1, the gain function of \( f_A^* \) can be written as:

\[
G_f(f_A^*) = \mathbb{E}_{(x,y) \sim P_D} \left[ \log \left( \frac{p_f(f(x))}{p_f(f(x)) + p_f'(f(x'))} \right) \right] \\
+ \mathbb{E}_{(x,y) \sim P_{D'}} \left[ \log(1 - \frac{p_f(f(x))}{p_f(f(x)) + p_f'(f(x'))}) \right] \\
= -\log(4) + 2 \cdot JS(p_f(f(x))|p_f'(f(x')))
\]

Where \( JS(p_f(f(x))|p_f'(f(x'))) \) is the Jensen–Shannon divergence between the two distributions. Since \( JS(p_f(f(x))|p_f'(f(x'))) \) is always non-negative and equals 0 if and only if \( p_f(f(x)) = p_f'(f(x')) \), the global minimum value that \( G_f(f_A^*) \) can possibly have is \(-\log(4)\) if and only if \( p_f(f(x)) = p_f'(f(x')) \) [Goodfellow et al., 2014]. This means that the prediction of classification model \( f \) for both the training set and non-training set has the same probability distribution. In this case, the classification model can be totally protected from MIA, the inference model can only flip a coin to make the determination with the possibility of 0.5. We use \( d \) to represent the Jensen–Shannon divergence \( JS(p_f(f(x))|p_f'(f(x'))) \) between the probability distributions of \( f \)'s outputs for the training set and non-training set. The larger \( d \) is, the higher the maximum gain of the inference model is. In other words, the more vulnerable the classification model is. Thus, any methods that can smaller \( d \) can have a defending effect against MIA. Intuitively, weight pruning can prevent over-fitting. Thus it will have a smaller \( d \).

5 Evaluation

In this section, we apply MIA-Pruning to different classification models with various DNN structures, mainly from two perspectives: the defense performance of our model and the computation cost benefit we obtain.

5.1 Classification Model

To evaluate our proposed method, we apply MIA-Pruning on different DNN models, including LeNet-5, VGG16, MobileNetV2, ResNet-18, on different datasets (e.g., MNIST, CIFAR-10, CIFAR-100, ImageNet). We use LeNet-5 on MNIST dataset, and VGG-16, MobileNetV2 and ResNet-18 to classify CIFAR-10 and CIFAR-100 dataset. We also use MobileNetV2 and ResNet-18 models on the ImageNet dataset to show the scalability of our proposed method.

We set the plain training of the classification model, i.e., without adversary training or MIA-Pruning, as our baseline. We compare the classification accuracy of the classification (target) model and the attack accuracy of the inference (adversary) model between MIA-Pruning, Min-Max Game, and baseline. We also compare MIA-Pruning with the popular method DP on CIFAR-10 dataset. We followed the same architecture with the four layers CNN classification model in [Rahman et al., 2018] and compare our results with the reported results in [Rahman et al., 2018]. For MNIST datasets, we use LeNet-5 as the classification model and implement DP with the noise parameter \( \epsilon \) as 6.28. The detailed setting of training can be found in Appendix Section 4.

Table 1: Comparison of classification accuracy and membership attack accuracy on different image datasets between model baseline, MIA-Pruning, and Min-Max Game.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>_baseline</th>
<th>MIA-Pruning</th>
<th>Min-Max Game</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>99.3%</td>
<td>67.00%</td>
<td>93.99%</td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>91.28%</td>
<td>61.99%</td>
<td>93.98%</td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>90.09%</td>
<td>62.75%</td>
<td>93.98%</td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>65.71%</td>
<td>66.15%</td>
<td>65.71%</td>
</tr>
<tr>
<td>CIFAR-100</td>
<td>64.84%</td>
<td>67.28%</td>
<td>65.71%</td>
</tr>
<tr>
<td>CIFAR-100</td>
<td>64.14%</td>
<td>66.15%</td>
<td>65.71%</td>
</tr>
<tr>
<td>MobileNetV2</td>
<td>73.6%</td>
<td>69.85%</td>
<td>65.71%</td>
</tr>
<tr>
<td>ResNet-18</td>
<td>71.4%</td>
<td>74.10%</td>
<td>65.71%</td>
</tr>
</tbody>
</table>

5.2 Inference Attack Model

To compare with the Min-Max game, we use the same neural network as the inference attack model as in [Nasr et al., 2018] for all experiments except CIFAR-10-CNN in Table 2. For CIFAR-10-CNN in Table 2, the inference attack model is the same as in [Rahman et al., 2018] (see Appendix Section 5).

5.3 Evaluation Results on MIA-Pruning

MNI, CIFAR-10 and CIFAR-100

We compare weight pruning (using MIA-Pruning) and Min-Max game to investigate if weight pruning can constrain the maximum gain \( G \) of the inference model, i.e., further reduce attack accuracy. We provide the attack accuracy and testing accuracy of baseline (without defense and pruning),
MIA-Pruning, and the Min-Max game as shown in Table 1. On MNIST, experimental results demonstrate that for LeNet-5, the attack accuracy using MIA-Pruning is 13.6% lower than the attack accuracy of baseline and is 2.6% lower than the attack accuracy of the Min-Max game. From the comparison between DP and MIA-Pruning shown in Table 2, MIA-Pruning achieves 25.13% lower attack accuracy than DP and with 2.66% higher testing accuracy of the classification model.

On CIFAR-10, the experimental results demonstrate that for VGG16, the attack accuracy using MIA-Pruning is 3% lower than the baseline attack accuracy and is 1.34% lower than the attack accuracy of the Min-Max Game. On the other hand, for MobileNetV2, the attack accuracy using MIA-Pruning is 3.77% lower than the baseline attack accuracy and is close to Min-Max game. As shown in Table 2, on a four-layers CNN [Rahman et al., 2018], MIA-Pruning has 1% lower attack accuracy than DP, while MIA-Pruning has 7.36% higher testing accuracy of the classification model than DP. On CIFAR-100, the experimental results demonstrate that for VGG16, the attack accuracy using MIA-Pruning is 9.1% lower than the baseline attack accuracy and is approximately 10% lower than the Min-Max game. On the other hand, for MobileNetV2, the attack accuracy using MIA-Pruning is 3.48% lower than the baseline attack accuracy and is close to the Min-Max Game.

The results indicate that using weight pruning can help against MIA, and weight pruning is more effective than using Min-Max Game. On the other hand, weight pruning has significantly less utility cost than DP. And our experiment also shows that DP is hard to achieve privacy-preserving with negligible utility loss. Also, base on the experiment in [Rahman et al., 2018], to achieve the same level of attack accuracy, the test accuracy under the DP method is under 70% in the best case, 25% in the worst case on CIFAR-10 by different noise parameter $\epsilon$. In addition, weight pruning brings another benefit shown in Table 4, i.e., we achieve 15.78X model size reduction for LeNet-5 on MNIST, at least 10.06X model size reduction for on CIFAR-10/CIFAR-100 among VGG16, MobileNetV2, and ResNet-18, which is extremely helpful for deploying DNNs on mobile devices. Figure 3 (a)-(c) show the weight distributions in different classification models from baseline, MIA-Pruning, and Min-Max game. We can observe that the weights after pruning are much less than the baseline model and Min-Max game model (both without pruning).

Next, we investigate classification loss of baseline (without pruning and defense), MIA-Pruning, and Min-Max Game. Taking CIFAR-10-VGG16 as an example, Figure 2 shows the classification loss of baseline, MIA-Pruning, and Min-Max Game, respectively, in the upper row. The classification loss of MIA-Pruning converges rapidly in less than 20 epochs. In addition, it has the highest final classification loss when the model is fully trained. In other words, MIA-Pruning prevents overfitting instead of reducing the classification loss on training data arbitrarily low. We train the membership inference model based on the predicted outputs of the well-trained classification model. We plot the testing accuracy of membership inference attack during the inference model training process.

<table>
<thead>
<tr>
<th>Model</th>
<th>Testing accuracy</th>
<th>Attack accuracy</th>
<th>Testing accuracy</th>
<th>Attack accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST-LeNet-5</td>
<td>96.64%</td>
<td>78.54%</td>
<td>99.30%</td>
<td>53.41%</td>
</tr>
<tr>
<td>CIFAR-10-CNN</td>
<td>68.10%</td>
<td>58.30%</td>
<td>75.46%</td>
<td>57.36%</td>
</tr>
</tbody>
</table>

Table 2: Comparison of classification accuracy and membership attack accuracy between DP and MIA-Pruning.

<table>
<thead>
<tr>
<th>Model</th>
<th>Classification accuracy</th>
<th>Membership accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>ImageNet-MobileNetV2</td>
<td>71.88%</td>
<td>66.90%</td>
</tr>
<tr>
<td>ImageNet-ResNet-18</td>
<td>69.76%</td>
<td>66.20%</td>
</tr>
</tbody>
</table>

Table 3: Comparison of classification accuracy and membership attack accuracy on ImageNet between model baseline and MIA-Pruning.

Figure 2: Upper row: the classification loss along with epochs during training classification model. Lower row: membership attack accuracy during training the inference model gain the trained classification model. Classification models baseline, MIA-Pruning, and Min-Max game are shown in (a-c) respectively.

Figure 3: Distribution of weights in VGG16 trained on CIFAR-10 for (a) baseline, (b) MIA-Pruning, (c) Min-Max Game, and (d) MIA-Pruning & Min-Max.
in the lower row in Figure 2. The adversary attack accuracy is measured by averaging the adversary’s correct determination percentage among all adversary determination for the observed data records [Nasr et al., 2018].

<table>
<thead>
<tr>
<th>Data</th>
<th>Model</th>
<th>Weights (#)</th>
<th>Weights after pruning (#)</th>
<th>Weights reduction ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>LeNet</td>
<td>60K M</td>
<td>15K M</td>
<td>0.25 x</td>
</tr>
<tr>
<td>CIFAR-10/100</td>
<td>ResNet-18</td>
<td>11.17 M</td>
<td>1.00 M</td>
<td>10.17 x</td>
</tr>
<tr>
<td></td>
<td>MobileNetV2</td>
<td>3.46 M</td>
<td>0.36 M</td>
<td>9.36 x</td>
</tr>
<tr>
<td>ImageNet</td>
<td>ResNet-18</td>
<td>11.17 M</td>
<td>1.04 M</td>
<td>10.13 x</td>
</tr>
<tr>
<td></td>
<td>MobileNetV2</td>
<td>3.46 M</td>
<td>0.36 M</td>
<td>9.36 x</td>
</tr>
</tbody>
</table>

Table 5: Comparison of classification accuracy and membership attack accuracy between MIA-Pruning and MIA-Pruning & Min-Max.

**ImageNet**

The experimental results for ImageNet are shown in Table 3, which demonstrates that for MobileNetV2, the attack accuracy using MIA-Pruning is 2.11% lower than the baseline, then, for ResNet-18, the attack accuracy using pruning is approximately 5% lower than the baseline. The weight reduction ratio is 3.37 \times for ResNet-18 and 3.27 \times for MobileNetV2 compared with the baseline weights.

**5.4 Evaluation Results on MIA-Pruning & Min-Max**

The experimental result for the MIA-Pruning & Min-Max is showed in Table 5. The experiment results demonstrate that for CIFAR-10-VGG16, the attack accuracy of Pruning & Min-Max is 54.67%, which is 3.03% lower than the attack accuracy of MIA-Pruning. And for CIFAR-100-VGG16, the attack accuracy of pruning & Min-Max is 57.65%, which is 1% lower than the attack accuracy of MIA-Pruning. For MNIST-LeNet-5, the attack accuracy of pruning & Min-Max is close to the attack accuracy of MIA-Pruning. Figure 3 (d) shows the distribution of weights in classification models from MIA-Pruning & Min-Max. We can also observe that after pruning, the weights are much less than the baseline model.

**5.5 MIA-Pruning Analysis**

In general, for the same type of model, the more overfitting the model is, the more vulnerable it is to MIA. The least generic the distribution of training data is, the more information it leaks. MIA-Pruning achieves parameter sparsity by pruning non-critical weights, thus can potentially reduce the overfitting caused by over parameterization. Taking CIFAR-10-VGG16 as an example, we compare the prediction on training data and non-training data between baseline, MIA-Pruning, and Min-Max Game. The baseline predicts the highest probability for the correct class in the training data. In other words, it has the highest prediction certainty, or it fits the training data best. However, the baseline has significantly lower prediction certainty on the testing data. The difference of prediction between the training data and testing data can be learned by the inference model to distinguish the membership. Since models with pruning or/and Min-Max Game reduce overfitting, they have lower prediction certainty on the training data, but similar to the prediction certainty on the testing data. The detail of pruning rate settings is in Appendix Section 6.

To summarize the difference of prediction between training and non-training data quantitatively, we show the MIA accuracy along with the difference of classification accuracy between training and non-training data for each class in CIFAR-10 in Table 6. The larger the training-non-training accuracy gap is, the higher the membership attack accuracy is.

6 Conclusion

In this work, we jointly formulate weight pruning and MIA as MIA-Pruning and provide an algorithm to solve the problem. We theoretically analyze and conclude that our proposed algorithm can protect the information privacy from MIA and achieves competitive accuracy with the original DNNs. And we evaluate our method on LeNet-5, VGG16, MobileNetV2, ResNet-18 on different datasets including MNIST, CIFAR-10, CIFAR-100, and ImageNet. From experimental results, we see weight pruning can significantly reduce the information leakage from MIA.

Our proposed method outperforms DP on MIA. Compared with our MIA-Pruning, our MIA-Pruning & Min-Max game can achieve the lowest attack accuracy so that it maximally enhance DNN model privacy. Thanks to the hardware-friendly characteristic of weight pruning (reducing weight storage and computational operation), our proposed MIA-Pruning is very helpful for deploying DNNs on mobile devices. We hope our proposed method will shed some light on the increasing membership privacy concerns when applying DNNs on user-sensitive data such as business and medical datasets on mobile devices.

**Acknowledgements**

This work has been supported in part by the following NSF grants: 1743418 and 1843025. This work originated from the course project of CSE5095 Advances in Deep Learning taught by Prof. Caiwen Ding. We also thank Dan Guo for many helpful discussions.
References


