Membership Inference Attack on Classification Models

Yang Yiliu
1155157082

Abstract

In this report, the author introduce the concept of Membership Inference Attack (MIA). The author also introduce different MIA methods including metric-based attacks, machine learning attacks. To stop MIA from stealing private data, some MIA defense methods are also developed such as Mix-up[18] and DP-SGD[16]. In the experiments part, the author tested some assumptions made by the original papers and evaluate different MIA and MIA defense methods. The author also discuss the relationship between overfitting and MIA, and MIAs on text generation models.

1 Introduction

Recently machine learning (ML) models have achieved remarkable results in different areas, such as face recognition, natural language processing and drug discovery. One key problem in training a successful ML model is the requirement for a large dataset. Depending on the source of data, a dataset can contain personal information. Since users will never want their personal information to be made public on the Internet, an ML model should not leak any individual information from the training dataset. However, [3] has shown that an ML model will leak individual information in some situations, which leads to privacy issues in both data collecting and model training procedures. The Membership Inference Attack (MIA) is one of the attacking methods aiming to classify whether the training dataset contains a certain record or not. To reduce the ability of a model to leak information, several methods [9] have been developed to defend against MIA.

The attack method of machine learning model attack can be categorized as two settings, i.e., white-box attack setting and black-box attack setting. The white-box attack is the setting that the attacker can access all information about the target model, including model architecture and trained parameters. The black-box attack is the setting that the attacker can only access limited information about the target model. [6]

Several methods have been developed in the MIA and defense field. For example, in the attacking field, random guessing, which is the strategy that randomly guess whether the record is in the training data or not, is a commonly used baseline method in the MIA, achieving an accuracy of 50%. Apart from the random guessing, these is also a commonly used baseline called Prediction Correctness based MIA [16], which infers a record as a member if it is correctly predicted by the target model, otherwise the attacker infers it as a non-member, achieving an accuracy of $\frac{1+g}{2}$[1]. In the defense field, regularization has been proved to be an effective method to defend the MIA due to its ability of reducing overfitting.

This project aims to re-implement several famous MIA methods to attack models trained in real-world datasets to test their attacking accuracy. This project also apply some defense methods to target models to evaluate their performances.

This report will not contain related work because this report will not cover new attack or defense method.

[1] g is the generalization gap, i.e., $\mathcal{L}_{train} - \mathcal{L}_{test}$. Detailed proof is in 2.1.

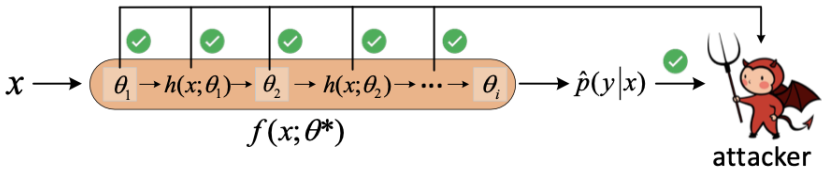


Figure 1: Overview of white-box attack

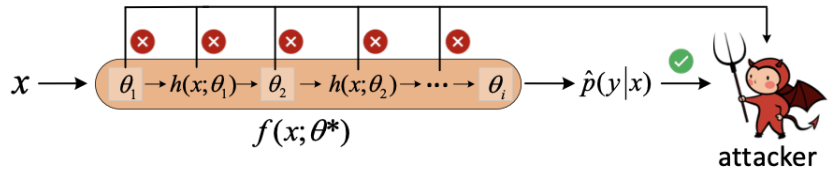


Figure 2: Overview of black-box attack

2 Attack

Theoretically, the Membership Inference Attack is defined as a binary classification problem by identifying whether a record is used as part of the training dataset of a given Machine Learning model. For simplicity, the machine learning model to be attacked is called the **Target Model**, and a record in the training dataset is called a **Record Data**. Since MIA is a binary classification problem, any metric of binary classification can be used to evaluate the effect of MIA. In this report, the author will mainly use the accuracy to evaluate because the testing dataset for MIA usually contains 50% member data and 50% non-member data.

Depending on what knowledge the attacking model has, the attack is divided into two cases.[6] The white-box attack is the setting that the attacker can access all information about the target model, including model architecture, intermediate model output and trained parameters. The black-box attack is the setting that the attacker can only access limited information about the target model, which are 1. The black-box access to the target model oracle. 2. The type of target model architecture and training data distribution.[12] In practice, unless a hacker have accessed to the server where the machine learning model located, an attacker cannot get the whole information about the target model. Therefore, the black-box attack setting is more realistic than the white-box attack setting. And in this report, the author will only focus on the black-box attack setting.

Generally, in the black-box attack setting, the attacker will have the information of the estimated model architecture with implemented shadow models $\mathcal{M}_{\text{shadow}}$ accordingly, the estimated training dataset (shadow dataset) $\mathcal{D}_{\text{shadow}}$. For a specific record, the attacker will further have the record feature x, the record label y and the model output $\hat{y}$ by querying the black-box target model oracle with x. The attacker will want to get the private training dataset $\mathcal{D}_{\text{private}}$

2.1 Baselines

Random guessing is the most trivial algorithm in the MIA field. The strategy of random guessing is to randomly guess whether the record is a member or not without considering the information of target model and the record. Since the MIA is a binary classification problem, the expected attack accuracy is 0.5.

Another trivial algorithm is the baseline attack (also called global-label attack). The strategy of baseline attack is to guess the record is a member if the model correctly predict the label of this record.

$$\mathcal{A}_{\text{baseline}}(y, \hat{y}) = \mathbb{1}(y = \text{argmax}\{\hat{y}\}) \tag{1}$$

Theorem 2.1 (Baseline Attack Accuracy)**.** *If the testing dataset has 50% member data and 50% non-member data, the attack accuracy of the baseline attack will be* $\frac{1+g}{2}$.

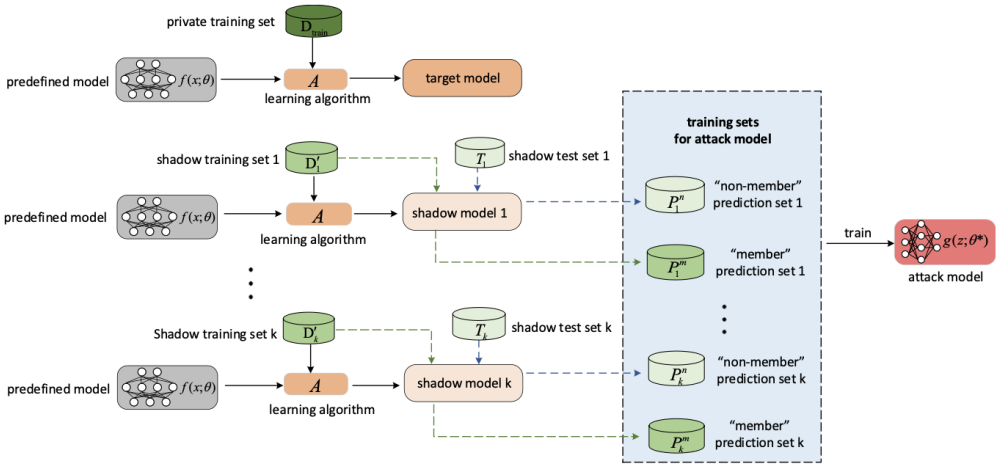


Figure 3: Overview of shadow training

In this theorem, g is the generalization gap, i.e., training accuracy minus testing accuracy in the classification task.[16]

Proof.

$$\begin{aligned}&\text{Accuracy}\\&= P(y \in \mathcal{D}_{\text{train}}.y)P(y = \text{argmax}\{\hat{y}\}|y \in \mathcal{D}_{\text{train}}.y) + P(y \notin \mathcal{D}_{\text{train}}.y)P(y \neq \text{argmax}\{\hat{y}\}|y \notin \mathcal{D}_{\text{train}}.y)\\&= 0.5\text{Acc}_{\text{train}} + 0.5(1 - \text{Acc}_{\text{test}}) = 0.5(1 + \text{Acc}_{\text{train}} - \text{Acc}_{\text{test}}) = \frac{1+g}{2}\end{aligned}$$

□

This theorem suggests a linear relationship between MIA accuracy and the generalization gap g. We will further test the relationship in the experiment part of this report.

2.2 Shadow Training

Training data collection is a significant problem in MIA model training because the attacker cannot collect any useful data from the target model. Shadow Training [12] is proposed to solve this problem. The idea of shadow training is an attacker can create multiple shadow models to mimic the behavior of the target model. The overview of shadow training is shown in the figure 3. For these shadow models, the attacker has their training datasets and test datasets. The attacker can train its model based on the data collected from shadow models. In detail, if there are k shadow models $\mathcal{M}_1, ..., \mathcal{M}_k$, the shadow dataset $\mathcal{D}_{\text{shadow}}$ will be divided into $2k$ disjoint datasets $\mathcal{D}_1, ..., \mathcal{D}_k$ and $\mathcal{D}'_1, ..., D'_k$. Then train the i-th shadow model $\mathcal{M}_i$ with $\mathcal{D}_i$. Finally pass $(\mathcal{D}_i, \mathcal{D}'_i)$ to $\mathcal{M}_i$ to generate member data and non-member data. With the help of generated member data and non-member data, the attacker can implement advanced attack methods.

2.3 Metric-based Attacks

There are many metrics can be used to evaluate a classification model output, which can also be used to determine the membership of a record. Metric-based attacks determine the membership of a record by first calculating metrics on their prediction vectors. The calculated metrics are then compared with a preset threshold τ to decide the membership status of the record. Based on different metric options, there are three major types of metric-based attacks, i.e., Loss-based MIA, Confidence-based MIA, and Entropy-based MIA.[6] The preset threshold can be set manually or according to the result of shadow training.

Loss-based MIA The attacker guess the record is a member if its loss value is lower than a threshold. The intuition is that the optimization objective of a classification model is the loss value, and the loss value of a training record should be lower than the loss value of a testing record.[16]

$$\mathcal{A}_{\text{loss}}(\hat{y}, y) = \mathbb{1}(\mathcal{L}(\hat{y}; y) \leq \tau). \tag{2}$$

where $\mathcal{L}(\cdot;\cdot)$ is the cross-entropy loss function.

Confidence-based MIA The attacker guess the record is a member if the greatest confidence score is greater than a threshold. The intuition is similar to the loss-based MIA, but it further considers that the optimum of the cross-entropy loss function is the one-hot vector of the ground-truth label, which means the maximum confidence score of a training member output should be greater than that of a testing member output.[11]

$$\mathcal{A}_{\text{conf}}(\hat{y}) = \mathbb{1}(\max\{\hat{y}\} \geq \tau). \tag{3}$$

Entropy-based MIA The attacker guess the record is a member if the Shannon entropy of its confidence scores is lower than a threshold. The intuition is also similar to the confidence-based MIA, but it uses Shannon entropy to evaluate, which means the Shannon entropy of a training member's confidence scores should be lower than that of a testing member's confidence scores.[11]

$$\mathcal{A}_{\text{entr}}(\hat{y}) = \mathbb{1}(H(\hat{y}) \leq \tau). \tag{4}$$

where $H(\cdot)$ is the Shannon entropy function.

2.4 Machine Learning Attack

Unlike metric-based attacks relying on the evaluation metrics and the preset threshold, machine learning attack requires less hypothesis of the model output, and it may learn complex relationship between members and non-members. But the machine learning attack is more complex and computational. machine learning attack determine the membership of a record by passing the prediction vector and the ground-truth label to a machine learning model. Then the machine learning model will binarily classify the membership of this record. The machine learning model can be trained using the data generated from the shadow training.

$$\mathcal{A}_{\text{ML}}(\hat{y}, y) = \phi(\hat{y}, y). \tag{5}$$

where ϕ is the trained binary classification machine learning model.

2.5 Similarity-based Attack

Instead of directly using shadow training, similarity-based attack aims to train multiple shadow models to capture the output difference of using the query record during attacking. In detail, for a given record feature x and label $\hat{y}$, the attacker creates k shadow datasets $\mathcal{D}_1, ..., \mathcal{D}_k$, and trains $2k$ shadow models $\mathcal{M}_1, ..., \mathcal{M}_k$ and $\mathcal{M}'_1, ..., \mathcal{M}'_k$, where $\mathcal{M}_i$ is trained with $\mathcal{D}_i$ and $\mathcal{M}'_i$ is trained with $\mathcal{D}_i \cup \{x\}$. Then the attacker can compare the similarity between $\mathcal{M}(x)$ and $\hat{y}$, and the similarity between $\mathcal{M}'(x)$ and $\hat{y}$ to determine the membership of this record. A typical choice of the similarity measurement is the KL Divergence.[10]

3 Defense

This section will introduce some defense methods against MIA.

3.1 Score Masking

Score masking the simplest way to defend MIA for classification models. The idea of score masking is to mask some unimportant confidence scores and only leave the top-K confidence scores. For example, in a classification task of 10 classes, the target model only provides the largest three

confidence scores. Then the attacker can only get limited information of this model, and the MIA accuracy will be reduced while keep the accuracy on the classification task.

Another kind of score masking is to adjust the prediction vector, adding crafted noise to the prediction vector or only output the predicted label without the confidence score. If the target model only output the predicted label, then the only method to MIA is the baseline attack.[6]

3.2 Regularization

Regularization is one kind of methods to reduce the overfitting degree of a machine learning model. As the baseline attack suggests, reducing overfitting degree will also reduce the MIA accuracy on this model. As a result, L_2 regularization, dropout[13], data augmentation, model stacking, early stopping[14] can be used to defend MIA.

Mix-up training augmentation[18] is another regularization method to reduce the overfitting degree. The idea of Mix-up is to combine two instances into one instance to train the machine learning model at every step.

$$\tilde{x} = \lambda x_i + (1-\lambda)x_j$$
$$\tilde{y} = \lambda y_i + (1-\lambda)y_j$$

where $\lambda \sim Beta(\alpha, \alpha)$ for $\alpha \in (0, \infty)$.

3.3 Differential Privacy

Differential privacy (DP) is a technique that preserves the privacy of individual data points within a dataset by introducing controlled noise. Using DP based method, private information is theoretically protected. Some DP based methods have been developed in the machine learning field. DP-SGD[1] is a DP-based machine learning method protecting privacy by adding Gaussian noise to the calculated gradient at every step. The detailed mathematical proof and implement method can be found in the original paper.

4 Experiments

In this experiments part, the author will show the results of self-implemented MIAs by comparing different MIA methods and testing the assumptions raised by original papers, including (1) comparison between metric-based attacks, (2) relationship between generalization gap and MIA accuracy, (3) transferability between different model architecture and different dataset, (4) effectiveness of MIA defense, and (5) accuracy influence after using defense method. The code is available at https://github.com/Yasgant/4010project. This part will not contain the results of similarity-based attack and DP-SGD because both of these require extremely huge computational power.[2]

For target models and shadow models, they are trained in the setting of `batch_size=128`, `optimizer=Adam`, `lr=1e-3`, `max_epochs=50`.

For the machine learning attack model, the author used MLP model with one hidden layer of size $(10, 20, 50, 70, 100, 150, 200)$ and 50 epochs to fine-tune the result.

4.1 Data

Samples of the following datasets are shwon in Figure 4.

MNIST MNIST[15] is a handwritten digit dataset containing 70,000 gray-scale 28x28 images. There are 10 classes of different digits in this dataset. This dataset serves as a easy dataset where almost all machine learning models will have low generalization gap.

[2]Similarity-based attack needs to train at least two ML models for each query. Training procedure using DP-SGD is always very slow and needs thousands of epochs to converge.

Figure 4: Samples of MNIST (left), CIFAR-10/100 (middle), A1K (right) datasets

CIFAR CIFAR-10 and CIFAR-100 datasets[7] are benchmark datasets used to evaluate image recognition algorithms. CIFAR-10 consists of 60,000 32x32 colorful images in 10 classes. CIFAR-100 is similar to CIFAR-10, but has 100 classes containing 600 images each. CIFAR datasets serve as appropriate datasets where some machine learning models have low generalization gap and some have high generalization gap.

A1-Kaggle A1-Kaggle (A1K)[3] is one of the assignment datasets of AIST4010/ESTR4140 course. This dataset consists of 95,134 64x64 colorful images in 1,000 classes. This dataset serve as a hard dataset where only carefully designed machine learning model can have low generalization gap and most of the models will have very high generalization gaps.

In order to do the shadow training, every dataset is equally divided into 4 parts $\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}}, \mathcal{D}'_{\text{train}}, \mathcal{D}'_{\text{test}}$. The target model is trained with $\mathcal{D}_{\text{train}}$ and the shadow model is trained with $\mathcal{D}'_{\text{train}}$. Then $\mathcal{D}'_{\text{train}}$ and $\mathcal{D}'_{\text{test}}$ are used to generate shadow data. $\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}}$ are used to evaluate the the accuracies of different MIA methods.

4.2 Model Architecture

MLP In this report, MLP architecture is one kind of machine learning models with only fully connected layers which are not carefully designed. The specific model architecture varies according to training dataset.

CNN In this report, CNN architecture is one kind of machine learning models with both fully connected layers and convolution layers, which are also not carefully designed. The specific model architecture also varies according to the training dataset.

Alexnet Alexnet[8] is a well-known deep learning model for image recognition. AlexNet gained popularity after winning the ImageNet competition in 2012, and it has since become a popular model for training on various datasets.

Resnet ResNet[4], short for Residual Network, is another deep learning model for image recognition. ResNet is famous because of its use of residual connections, which help improve the flow of information through the network. This design allows ResNet to be much deeper than other CNNs while still maintaining high performance, making it a popular choice for training on diverse datasets.

4.3 Metric-based Attacks

We tested different MIAs on different datasets and target model architectures, where target model and shadow model shares a same architecture. Table **??** and table 4.3 show, loss-based attack always has the greatest AUC score in metric-based attacks. We can conclude that loss value is the best metric to determine the membership.

This result is understandable because the optimization objective of a classification problem is to minimize the loss value. During training, the loss value of training instances are reduced more than testing instances. Then directly use loss value as the metric to determine the membership should be a good method.

[3] https://www.kaggle.com/competitions/aist4010-spring2023-a1

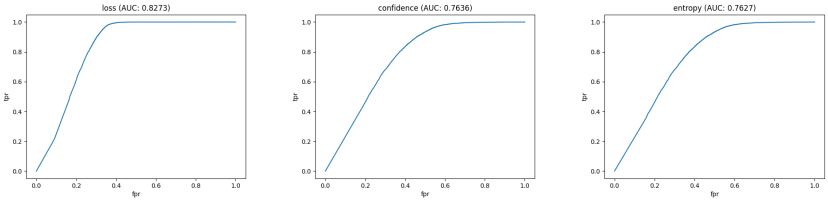
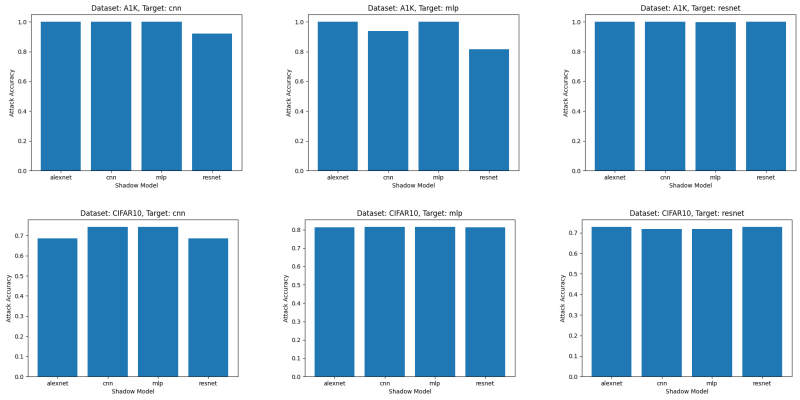


Figure 5: ROC Curves of Metric-based MIAs on CIFAR-10 dataset



4.4 MIA Accuracy vs Generalization Gap

We tested different MIAs on different datasets and target model architectures, where target model and shadow model shares a same architecture. In figure **??**, we can find that there is a rough linear relationship between MIA accuracy and generalization gap.[4] As the generalization gap can reveal the overfitting degree of a model, this shows there is a relationship between MIA and overfitting. Detailed discussion about MIA and overfitting will be in the section 5.1.

4.5 Transferability

Shadow models[12] are believed to be able to mimic the behavior of target model. We tested different MIAs on different datasets and target model architectures, where target models have different architecture than shadow models.

Different Model Architectures In this setting, the shadow dataset has a same distribution with the target model. Figure **??** shows there is only a small difference between different model architectures.

Different Data Distributions In this setting, the shadow dataset has a different distribution than the target model. Figure **??** shows there is also only a small difference between different distribution.

4.6 MIA Defense

We tested different MIA defense methods on different MIA methods, different model architecture and dataset. As the result in 4.5, we only tested the situation where target model and shadow model share the same architecture and data distribution.

In figure **??**

In figure **??**,

[4]Entropy-based attack is unstable in various datasets and model architecture. This may be because the numerical unstable of entropy calculation. Also this is the reason why entropy-based attack is a bad MIA method.

5 Discussions

5.1 Ideas of MIA

Many papers have pointed out that overfitting of the target machine learning model is the key factor for the MIAs. The overfitting in a machine learning model describes that the model behaves better when encountering the training data than the testing data, i.e., it cannot generalize well on general data. The overfitting is usually caused by two reasons. One is the high model complexity and another is the small size of training data. Deep learning models are often overparameterized with high complexity, which allows them to memorize the training data. [17] Besides, the modern machine learning models are always trained using multiple epochs, which forces model to encounter the same data multiple times. When a model has memorized the training data, it will have a high probability to behave differently on training data and fail to generalize to testing data.

There are also some theoretical guarantee of MIA. For example, the theorem in [2] shows that there exists a MIA method with accuracy more than 50% if the target model overfits the training dataset.

Theorem 5.1. *Assume a model has a generalization gap $g > 0$ and q of the evaluation dataset are member data. Then there exists a MIA method with the accuracy:*

$$\begin{aligned} \textit{Accuracy} &\geq \max\{q, 1-q, q\textit{Acc}_{\textit{train}} + (1-q)(1-\textit{Acc}_{\textit{test}})\} \\ &\geq \max\{q, 1-q, \min\{q, 1-q\}(1+g)\} \\ &> 1/2 \end{aligned}$$

This serves as a lower bound of the MIA accuracy. However, to my best, currently there is no theoretical guarantee of the lower bound of other MIA accuracy such as metric-based MIA and machine learning MIA.

Based on the analysis, reducing overfitting should be a good method to defend MIA, such as regularization disscussed before. However, there is a trade-off between the privacy protection and model effectiveness in the machine learning field as many data privacy papers suggest.[1]

The training dataset diversity is also another factor of MIA. If the training dataset is lack of diversity, then the machine learning model cannot generalize well. And thus this model will have a high probability to be attacked by MIA.

5.2 MIA on Large Language Models

MIA can also be applied on GANs, embedding models, and regression models.[6] But in this section, I will only talk about the MIA on text generation models. To my best, MIA on text generation models is still an open problem and the current SOTA model can only achieve roughly 51% accuracy on BERT.[5]

However, current large language models have a huge amount of parameters. For example, GPT-3 has 175 billion parameters while GPT-4 is considered to have more parameters than GPT-3. It is believable that these large language models suffer overfitting terribly. According to the disscussion in 5.1, if effective MIA can be applied on these models, many private data will be leaked to the Internet. When developers train these large language models, they should pay more attention on the privacy protection including API limitation and data masking.

6 Conclusion

In this project, the author introduced different membership inference attack methods and different MIA defense methods. In the experiment part, the author showed that MIA is effective on different model architectures and different datasets. The author also showed that defense methods reduce MIA accuracy to varying degrees without influencing the model accuracy. In the future, the author expect to further study more MIA methods not only on classification models and develop an effective MIA method on text generation models.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[2] Jason W Bentley, Daniel Gibney, Gary Hoppenworth, and Sumit Kumar Jha. Quantifying membership inference vulnerability via generalization gap and other model metrics. *arXiv preprint arXiv:2009.05669*, 2020.

[3] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *USENIX Security Symposium*, volume 6, 2021.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] Sorami Hisamoto, Matt Post, and Kevin Duh. Membership inference attacks on sequence-to-sequence models: Is my data in your machine translation system? *Transactions of the Association for Computational Linguistics*, 8:49–63, 2020.

[6] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37, 2022.

[7] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research).

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[9] Jiacheng Li, Ninghui Li, and Bruno Ribeiro. Membership inference attacks and defenses in classification models. In *Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy*, pages 5–16, 2021.

[10] Yunhui Long, Vincent Bindschaedler, and Carl A Gunter. Towards measuring membership privacy. *arXiv preprint arXiv:1712.09136*, 2017.

[11] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246*, 2018.

[12] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

[13] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[14] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[15] lecun yann. The mnist database.

[16] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE, 2018.

[17] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

[18] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.