M^4I: Multi-modal Models Membership Inference

Anonymous Author(s)
Affiliation
Address
email

Abstract

With the development of machine learning techniques, the attention of research has been moved from single-modal learning to multi-modal learning, as real-world data exist in the form of different modalities. However, multi-modal models often carry more information than single-modal models and they are usually applied in sensitive scenarios, such as medical report generation or disease identification. Compared with the existing membership inference against machine learning classifiers, we focus on the problem that the input and output of the multi-modal models are in different modalities, such as image captioning. This work studies the privacy leakage of multi-modal models through the lens of membership inference attack, a process of determining whether a data record involves in the model training process or not. To achieve this, we propose Multi-modal Models Membership Inference (M^4I) with two attack methods to infer the membership status, named metric-based (MB) M^4I and feature-based (FB) M^4I, respectively. More specifically, MB M^4I adopts similarity metrics while attacking to infer target data membership. FB M^4I uses a pre-trained shadow multi-modal feature extractor to achieve the purpose of data inference attack by comparing the similarities from extracted input and output features. Extensive experimental results show that both attack methods can achieve strong performances. Respectively, 72.5% and 94.83% of attack success rates on average can be obtained under unrestricted scenarios. Moreover, we evaluate multiple defense mechanisms against our attacks. The source code of M^4I attacks is publicly available at https://github.com/MultimodalMI/Multimodal-membership-inference.git.

1 Introduction

Machine learning has witnessed great progress during recent years and now has been applied to many multi-modal applications (models that can process and relate information from multiple modalities, such as image, audio, and text), such as audio-visual speech recognition [1], speech recognition [1], event detection [2], and image captioning [3–13]. However, prior research [14] has revealed that machine learning models are vulnerable to membership inference attack. Leaking membership status (*i.e.*, member or non-member of a dataset) in a multi-modal application leads to serious security and privacy issues. For example, if a person, unfortunately, gets a tumor and his medical records were used to train an image captioning model (*e.g.*, to predict whether the tumor is benign or malignant), the attacker can infer that the person has a tumor. Furthermore, in a large-scale neural language model, the attacker can infer detailed private data with the membership inference attack, such as address or phone number, from the training data by text sequence generation [15]. This work studies the privacy leakage of multi-modal models through the lens of membership inference attack.

1.1 Background

Membership inference. Previous studies [14, 16–29] on membership inference mainly target the classification models trained from scratch, and many of them focus on exploiting the confidence

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Table 1: An overview of membership inference research.

Target tasks	**Attacks**	**Shadow Model**	**Detailed Model Prediction** (*e.g.*, confidence vector)	**Final Model Prediction** (*e.g.*, label)	**Other Useful Information**
Classification	[14, 20, 31–38]	• or -	•	•	Model weights
Generation models	[39–45]	• or -	-	•	-
NLP	[15, 30]	•	• or -	•	Reference output
Embedding models	[16, 46–48]	•	• or -	• or -	Inner layer output/feature
Federated models	[25, 49–53]	• or -	• or -	• or -	-
Recommendation system	[54]	•	-	•	-
Multi-modal models (ours)	Metric-based attack	•	-	•	Reference output
	Feature-based attack	•	-	•	-

"•": the adversary needs the knowledge; "-": the knowledge is unnecessary.

scores returned by the target modal as shown in Table 1. Since the concept of membership inference in machine learning was established in Shokri *et al.*'s work [14], later approaches [18, 20, 30] further improved these methods by exploring different assumptions (multiple shadow models, knowledge of the training data distribution and model structure [16, 31]) about the attacker and by broadening the scenarios of membership inference attacks [14, 20]. In addition to classifier-based approaches [14, 20, 31–38], membership inference has been shown effective under multiple machine learning settings, such as generative models [39–45], natural language models [15, 30], embedding models [16, 46–48], and federated learning [25, 49–53]. However, to the best of our knowledge, to date there is no membership inference attack targeting at multi-modal models.

Multi-modal machine learning. Unlike single-modal machine learning, multi-modal machine learning aims to build models that can manage and relate information from two or more modalities [**?**]. A new category of multi-modal applications has been emerging, such as image captioning [3–13] and media description and generation [55–57].

In this research, we conduct membership inference on a multi-modal task, image captioning [58] that is the process of generating a textual description of an image [3–13]. In image captioning, the datasets consist of image-text pairs.

1.2 Our Work

Unlike the single-modal learning task, training a multi-modal model can make use of more sensitive data from different modalities. For instance, in the medical domain, medical report generation [59, 60, 58, 61–63] can reduce the workload of medical professionals and speed up the diagnosis process by automatically analyzing medical examinations, such as computed tomography and X-ray photographs. However, the existing membership inference models can hardly be applied to multi-modal models for the following two reasons: (*i*) the multi-modal models, cannot provide much information such as confidence scores of the model prediction. (*ii*) Due to the requirements of handling multi-modal data, many existing membership inference methods, which target single-modal learning models such as image or text classification, are difficult to be squarely applied to multi-modal models. This motivates us to investigate a new category of membership inference attack on multi-modal models, which is, given access to a multi-modal data and model, an attacker can tell if a specific data sample is adopted in the model training process. In this work, we propose **m**ulti-**m**odal **m**odels **m**embership **i**nference (M^4I). Our main contributions are as follows:

- To the best of our knowledge, we are the first to investigate membership inference attacks on multi-modal models. To achieve this, we propose a metric-based attack and a feature-based attack to conduct M^4I under different assumptions.
- We conduct extensive experiments for general image captioning on MSCOCO, FLICKR 8k, and IAPR TC-12 datasets and we receive promising results. Furthermore, we apply our M^4I to a medical report generation model to witness a similar performance trend in the real world.
- We perform a detailed analysis of possible defense strategies, which are differential privacy, l_2 regularization, and data augmentation. These approaches provide avenues for future design of privacy mitigation strategies against M^4I.

2 Problem Formulation

In this section, we introduce the membership inference attack and formulate the threat model.

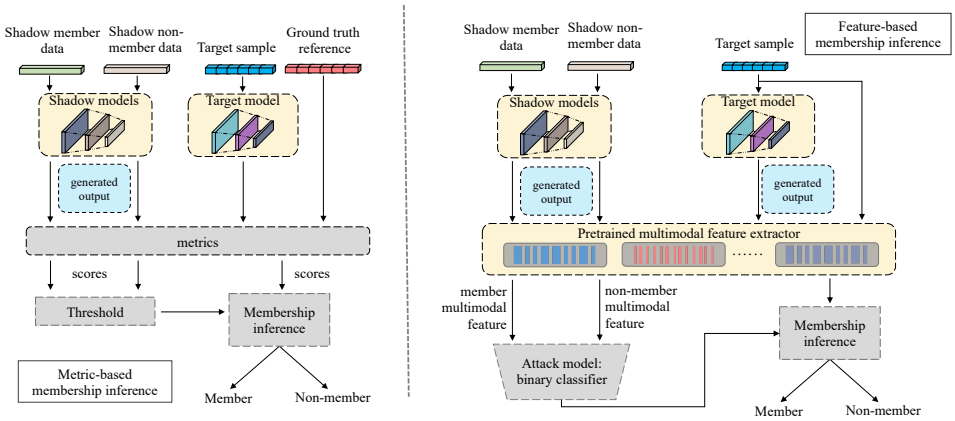


Figure 1: An overview of MB M^{4}I (left) and FB M^{4}I (right). The MB M^{4}I infers membership through similarity metric scores, while the FB M^{4}I utilizes a multi-modal feature extractor to distinguish member and non-member data on the feature level.

2.1 Membership Inference

Membership inference is a binary classification task. It aims to infer whether a data sample belongs to the training set of the target machine learning model. By querying the target model with a data sample, we can obtain the corresponding output. Then, the attack model takes it as the input and gives a binary output, indicating if this sample has been seen in the training set.

2.2 Threat Model

Attacker's goal. The adversary aims to infer whether an input data sample comes from the training dataset of a target image captioning model or not. If this input is in the target model's training dataset, we call it a member; otherwise, we call it a non-member. The adversary's goal is to accurately infer the member status (member/non-member of the target model's training dataset.)

Attacker's background knowledge. In our scenario, the attacker can query the target image captioning model, which provides lines of words as output without confidence scores. We consider this as the most general and difficult scenario for the attacker. In this research, we take training data distribution, model architecture, and ground truth reference into consideration as the attacker's other background knowledge. In particular, if an adversary knows the distribution, we assume the adversary has access to a shadow dataset with the same distribution as the target model's training set. With knowledge of model architecture, the adversary can use the same architecture for shadow model training. When the adversary does not know the target model's architecture, the adversary can assume another architecture with the same functionality for shadow model training. If the attacker knows the ground truth reference, we assume the attacker can use the text from the image-text pair dataset as ground truth. This knowledge will be utilized in metric-based membership inference. Depending on the difficulty to access the information for an attacker, we explore three different scenarios under different assumptions

- The **unrestricted** scenario, where an attacker has knowledge of both training data distribution and model architecture. This scenario is considered as a gray-box attack.
- The **data-only** scenario, when only the training data distribution is accessible for an attacker. This scenario is considered as a gray-box attack.
- The **constrained** scenario that neither knowledge of training data distribution nor model architecture is available to an attacker. This scenario is considered as a black-box attack.

3 Methodology

In this section, we propose the methodology of metric-based and feature-based M^{4}I attacks.

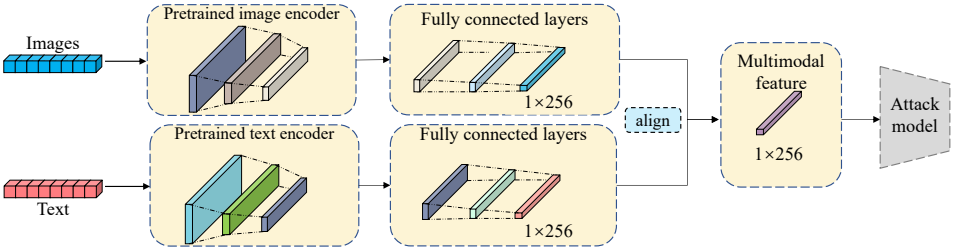


Figure 2: The multi-modal feature extractor is based on pretrained image encoder and text encoder.

3.1 Metric-based M^{4}I

The key intuition for the metric-based attack is that the output of a multi-modal model with an input sampled from its training dataset should be highly similar to or same as the ground truth reference of this sample. In the cases for image-text models such as image captioning or medical report generation, many similarity metrics such as Recall-Oriented Understudy for Gisting Evaluation (ROUGE) score [64] or Bilingual evaluation understudy (BLEU) score [65] are used to evaluate the similarity of the ground truth description and the output text. Here, we focus on the distinguishability of the membership information by these metrics, *i.e.*, the member and non-member data samples achieving differently on scores. Then we can leverage the shadow model to launch a metric-based membership inference attack. The metric-based membership inference (MB M^{4}I) can be divided into three main stages, namely metric calculation, preset threshold selection, and membership inference as shown in Figure 1.

Metrics calculation. Many metrics can be used to evaluate the similarity of the model output and corresponding ground truth. Here we utilize the ROUGE score [64] and BLEU score [65] for calculating the text-similarity for the image captioning model (see more details in Supplementary Materials).

Preset threshold selection. As the attacker cannot get the membership information of the training data, they can preset the threshold by training a shadow model $\mathcal{M}'$. With knowledge of the targeted task, the attacker can use a shadow dataset built up with collected public data to train a shadow model with a selected architecture. The attacker can split their shadow dataset into a member dataset and a non-member dataset and train the shadow model $\mathcal{M}'$ with the member dataset. By scoring the data from the member dataset and the non-member dataset, the adversary can easily get a threshold if there is only one metric applied. With more than one metric, the adversary can use a binary classifier to distinguish between the scores of member data and non-member data.

Membership inference. Based on the preset threshold, the attacker can build a threshold-based binary classifier $\mathbf{f}$ as an attack model. Then the adversary can send the target sample to the target model and get similarity scores of the output and ground truth reference by chosen metrics. The attack model can categorize the target sample as a member or a non-member.

3.2 Feature-based M^{4}I

Considering that the metric-based membership inference highly relies on the similarity metrics and the reference captioning of the target image, the reliability of the similarity metric can be limited in an image captioning task. Specifically, with an image of a cat and a ground truth description, "this is a cat sitting on the road", which the model has never seen, the model might output "there is a dog on the road." However, neither the BLUE nor the ROUGE metric can give this output a high score (BLEU-1: 0.625 and ROUGE-L: 0.66) and this image can be misidentified as a member by metric-based membership inference. Additionally, as mentioned in the threat model, the adversary may not have the knowledge of the ground truth reference. To solve the drawbacks of the metric-based membership inference, we propose feature-based membership inference.

The multi-modality feature extractor is able to learn the connection of meaningful information of image and text pairs [66]. The attacker can utilize the connection of the features of the input image and the output text for the membership inference.The feature-based membership inference can be divided into three main stages: multi-modal feature extractor training, attack model training, and membership inference, as shown in Figure 1 (right).

Multi-modal feature extractor training. Borrowed from [66], we build a multi-modal feature extractor with two different modal encoders and a layer to align features, as shown in Figure 2. With some public data or a published large-scale pretrained model, the attacker can easily get a pretrained image encoder f_i and a text encoder f_t. The encoders can be connected with a fully-connected layer so that the features of the image and text can be adjusted to the same shape. Then the attacker just needs to fine-tune the added layer by shadow dataset so that the features from the image encoder and text encoder can be aligned in the feature space.

For the ith sample(image $\mathbf{i}_i$ and text $\mathbf{t}_i$ pair), the feature can be formulated as follows:

$$\mathbf{F}_i = f_i(i_i); \mathbf{F}_\mathrm{t} = \mathrm{f_t}(\mathrm{t_i}), \quad \mathbf{z}_i = \mathbf{F}_i - \mathbf{F}_t, \tag{1}$$

where $\mathbf{F}_i$ and $\mathbf{F}_t$ represent the image feature and text feature, respectively. $\mathbf{z}_i$ represents the difference vector between multi-modal features of the image and text.

Parameter Optimization. Here we use stochastic gradient descent to update parameters in the multi-modal feature extractor (MFE), aiming to minimize the Euclidean distances between image features and text features:

$$L_{MFE} = \frac{1}{N}\sum_{i=1}^{N} ||\mathbf{z}_i||_2^2, \tag{2}$$

where $|| \cdot ||_2$ denotes L_2 norm and N is the number of training samples.

The goal of this loss function is to force the image encoder and text encoder to learn the corresponding feature representation from the image-text pairs. For each pair of images and texts, the multi-modal feature extractor would reduce the distance of their features in the same feature space.

Attack model training. An adversary can build an attack dataset of member multi-modal features and non-member multi-modal features using a shadow model. Those multi-modal features from the shadow member dataset can be labeled as "member" and those from the shadow non-member dataset can be labeled as "nonmember". With the attack dataset, the adversary can establish a binary classifier as the attack model $\mathcal{M}_{attack}$. The input of $\mathcal{M}_{attack}$ is $\mathbf{z}_i$ and the output of $\mathcal{M}_{attack}$ is a probability value between 0 and 1 for the membership status (0 means the attack model $\mathbf{z}_i$ is a non-member data and 1 means a member). For the ith sample, the prediction can be formulated as the following function:

$$\mathbf{y}_i = \mathcal{M}_{attack}(\mathbf{z}_i), \tag{3}$$

where $\mathbf{z}_i$ is the input of $\mathcal{M}_{attack}$ as well as the ith sample's multi-modal feature vector in our attack.

Membership inference. Test data i_{test} for the attack model will be sent to the target model to obtain the generated captioning t_{test}. Then we send the i_{test} and t_{test} pair to the multi-modal feature extractor to get the multi-modal feature $\mathbf{z}_{test}$. The trained attack model $\mathcal{M}_{attack}$ conducts a prediction given this multi-modal feature vector $\mathbf{z}_{test}$, i.e., $\mathcal{M}_{attack}(\mathbf{z}_{test}) = \mathbf{y}_{test}$, where $\mathbf{y}_{test}$ is a value indicating the probability that the test image belongs to members. According to the predicted results, the adversary infers the membership status of the test image. Specifically, when $\mathbf{y}_{test} > 0.5$, the test image is predicted to be a member. Otherwise, it is predicted to be a non-member.

4 Experiments

In this section, we evaluate our metric-based and feature-based M^{4}Is on image captioning models pre-trained on different image-text datasets.

4.1 Experiment Setup

Target model architectures. The target model is an image captioning machine learning model with an encoder-decoder architecture [3]. It is used to convert a given input image I into a natural language description T. This architecture involves using CNN layers (encoder) for feature extraction on input data, and LSTM layers (decoder) to perform sequence prediction on the feature vectors. In this work, we use Resnet-152[67] and VGG-16 [68] as heterogeneous encoders for different target model architectures.

Datasets. Our experiments are conducted on three different datasets: MSCOCO [69], FLICKR8k [70], and IAPR TC-12 [71], as detailed in Supplementary Materials.

Training target model. Our training method follows the settings in the work [3], excluding encoder architecture. We randomly sample 3,000 image-text pairs from each dataset as the "member dataset" to train the target image captioning model. Then, we randomly sample 3,000 image-text pairs from the rest of the datasets as the ground truth "non-member dataset" of the target model. Therefore, for each target image captioning model, we have 3,000 ground truth members and 3,000 ground truth non-members. We use Resnet-152 as well as VGG-16, with LSTM as the architecture for the target models.

Training shadow models. For both proposed M^4I attack methods, shadow models are indispensable. In the scenario where the attacker knows the target models' training data distribution, we randomly sample 6,000 image-text pairs from the corresponding dataset as the shadow dataset. In the scenario where the attacker does not have the knowledge of the data distribution, we randomly sample 6,000 image-text pairs from the other dataset. For instance, if the target model is trained with FLICKR8k dataset, we randomly sample 6,000 image-text pairs from the IAPR-TC12 dataset With these 6,000 image-text pairs, the attacker can split them into two 3,000 image-text datasets, which can be defined as the shadow member dataset and the shadow non-member dataset. The shadow member dataset will be used to train the shadow model and the shadow non-member dataset will be used in membership inference for attack model training.

Notations. To clarify the experimental settings, we use 2-letter and 4-letter combinations. For the 2-letter combinations, the first letter, *i.e.*, "C," "F" or "I", indicates the shadow (or target) dataset(which are MSCOCO, Flickr 8k, and IAPR dataset), and the second letter, *i.e.*, "R" and "V", indicates the training algorithm (which are Resnet-LSTM and VGG-LSTM). For example, "CR" is the combination of "MSCOCO+Resnet-LSTM". For the 4-letter combinations, the first two letters represent the dataset and algorithm of the shadow model and the last two letters denote the dataset and algorithm of the target model. For instance, "CRFV" means that the adversary establishes a shadow model with Resnet-LSTM on the MSCOCO dataset to attack a target model implemented by VGG-LSTM on the FLICKR 8K dataset, which is in the **constrained** scenario.

4.2 Metric-based M^4I Attack Performance

Implementation details. To evaluate the image captioning performance of our target model, we apply BLEU scores (BLEU-1, BLEU-2, and BLUE-3) and ROUGE-L score as the metric, which measures the similarity between the generated text and reference text. We establish a support vector machine (SVM) as the attack model that can divide the training data and non-member data. We combine the four scores calculated by the metrics as an input vector and send it to the attack model. When these scores are further away from the maximum-margin hyperplane computed by the SVM, the more confidence our attack model can have in classifying it as a member or non-member data point.

Unrestricted scenario. In this scenario, the target model's dataset distribution and algorithm are available, which are the maximum amount of knowledge that the adversary can gain from the target model. The complete results are shown in Figure 3, in which we compare our attack with Random Guess. Through the attack results, we can conclude that with the maximum amount of knowledge, the attacker can infer to great extent the membership by the scores from metrics. The overall accuracy is satisfactory that the highest accuracy (81%) is achieved when the target model is trained with the IAPR dataset.

In general, under the assumption that the adversary knows the algorithm and dataset distribution of the target model, our attack is fairly strong. But there are some cases that the attack is not effective when the target model is trained with the MSCOCO dataset. There are several reasons for the ineffectiveness. The first reason is that both metrics (BLEU and ROUGE) are direct comparisons of the generated text and reference text. They cannot distinguish the difference between meaningful content and some fixed sentence patterns. For example, the image captioning model can easily generate some meaningless sentences with a fixed pattern like "This is a picture of ...". In such a case, the metrics can still give meaningless output high scores, which can cause a high false-positive rate.

Data-only scenario. In this assumption, the attacker can only access a shadow dataset that has the same distribution as the training dataset. The experiment results are shown in Figure 3. It is shown in

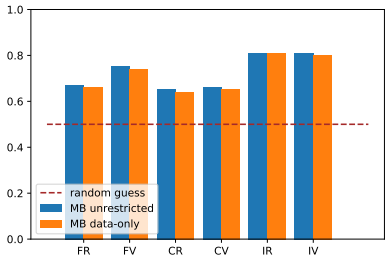


Figure 3: The metric-based attack success rate in **unrestricted** and **data-only** scenarios.

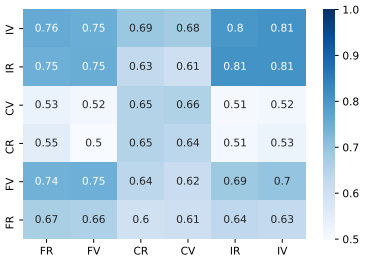


Figure 4: The Metric-Based Attack accuracy in all the scenarios.

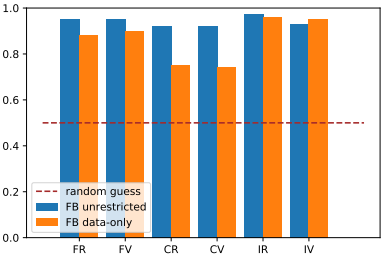


Figure 5: The feature-based attack success rate in **unrestricted** scenario and **data-only** scenario.

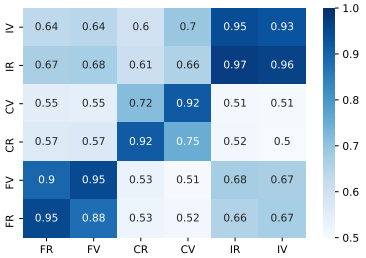


Figure 6: The feature-based attack performance in all the scenarios.

the results that in this scenario, the attack performance of metric-based membership inference drops as expected but only has a minor decay compared with the **unrestricted** scenario.

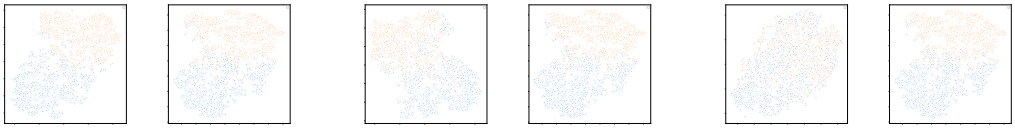
Constrained scenario. To this end, we discuss the final and hardest attack scenario where the attacker has no knowledge about the training dataset distribution or architecture. All the experiment results are shown in Figure 4. The x-axis indicates the shadow model's dataset and algorithms, and similarly the y-axis represents the target model's dataset and algorithms. The color bar shows the attack success rate. Note that all the results in **unrestricted** scenario and **data-only** scenario are also listed on this heat map. Under such a minimum assumption, the performance of metric-based attacks drops as expected. When the attacker does not know the model architecture and dataset information, the attack success rate of membership inference is 51% - 76%. We can find that attack performance should be highly related to the knowledge of the data distribution in the training set. As each line shows, the highest inference accuracy is always there when the attacker builds a shadow model with the data distribution same as the target model.

4.3 Feature-based M^{4}I Attack Performance

Implementation details. In our experiment, multi-modal feature extractors have been trained with the public data mentioned in Section 4.1. This feature extractor is a combination of an image encoder and a text encoder as shown in Figure 2. The attacker uses a two-layer multiplayer perceptron(MLP) as a binary classification attack model to infer the membership status. We provide more details of model architectures in Supplementary Materials.

Unrestricted scenario. The three assumptions are in the same condition in Section 4.2. For the unrestricted scenario, the results are shown in Figure 5. From the results, we can conclude that with knowledge of the data distribution and the architecture of the target model, the overall attack performance is better than the metric-based attack. Then, we visualize the data points (the multi-modal features of members and non-members) in two-dimensional space by t-distributed Stochastic Neighbor Embedding (t-SNE) [72]. Figure 7 shows the distributions of multi-modal features extracted by MFE from inputs and outputs of the target model and shadow model.

Without knowledge of the model architecture, attack performance drops as expected. However, in comparison with Figure 5, we can find that the accuracy decrease is even more than that in metric-



FR_shadow and FR_target FV_shadow and FR_target CV_shadow and FR_target

Figure 7: Visualization results by t-SNE in **unrestricted**, **data-only**, and **constrained** scenario, where blue points represent members and orange points represent non-members.

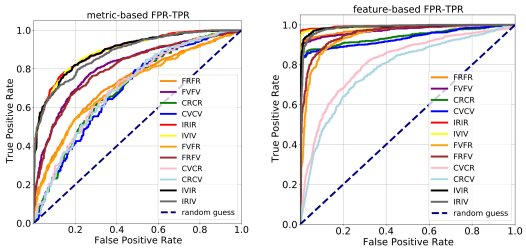


Figure 8: The true positive rate versus the false positive rate of metric-based membership inference (left) and feature-based membership inference (right). Here is the list of the **unrestricted** scenario, *i.e.*, "FRFR", and the **data-only** scenario, *i.e.*, "FVFR".

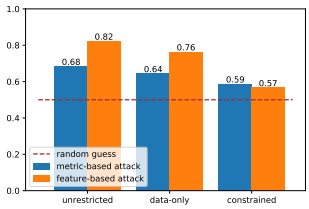


Figure 9: The metric-based attack and feature-based attack success rate on the medical report generation model. The x-axis indicates the attacker's background knowledge.

based attack's performance. As can be seen from Figure 7, the multi-modal feature distributions behave slightly different when the shadow model and target model are trained with different architectures. This difference can be hardly detected by a metric-based attack, as the scores from models trained with different architectures are quite similar. In comparison with direct similarity calculations, the feature extractor is able to find and emphasize some inconspicuous but important differences.

Constrained scenario. As the hardest scenario for attackers, our attack can still achieve good performance in some cases. The experiment results are shown in Figure 6. The x-axis indicates the shadow model's dataset and algorithms, and similarly, the y-axis represents the target model's dataset and algorithms. The color bar shows the attack success rate. The attack success rate of membership inference is 50% - 68%. As shown in Figure 7, without any information, the multi-modal features of the shadow model can only have limited overlapping with the target model. Comparing with the MB attack, FB attack relies more on the knowledge of data distribution.

4.4 Attack Performance with False-Positive Rates

Recent works [73] and [74] mentioned that it is insufficient to use only accuracy to show the performance of membership inference attack. Here we provide the receiver operating characteristic curve (ROC curve) to show the true positive rate (TPR) versus the false positive rate (FPR) for the **unrestricted** scenario and the **data-only** scenario. From the result shown in Figure 8, it can be concluded that the feature-based attack behaves better than the metric-based attack, which means while the feature-based attack correctly picks the member data from the attacker's dataset, this method also avoids categorizing non-member data into member data.

4.5 Attack Performance on Medical Report Generation

To evaluate our attack, we apply our methods to the Cross-modal Memory Networks for Radiology Report Generation (R2GenCMN) [75], which is a medical report generation model by taking chest X-Ray images as input and exporting medical reports.[1] This model proposes a method with a cross-modal mapping while generating the report. We evaluate the attack performance under the **restricted** scenario, **data-only** scenario, and **constrained** scenario. Experiment setting can be seen in Supplementary Materials. The experiment results are shown in Figure 9. We find that for the medical

[1] We directly run the model with the open-source code provided at https://github.com/cuhksz-nlp/r2gencmn.

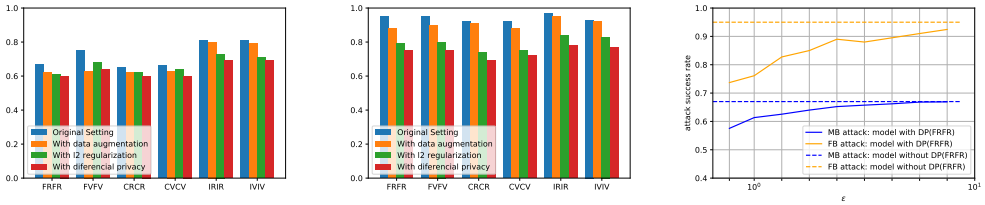


Figure 10: Comparison of attack success rate before and after applying three countermeasures. The left figure shows the results for metric-based attack and the middle figure shows the results for feature-based attack. The right figure shows the attack success rate for differential privacy training with different ϵ, given $\delta = 10^{-5}$.

report generation real-world task, our attack methods can still achieve strong performance. When the attacker has knowledge about both the target model's architecture and training data distribution, the feature-based attack accuracy can achieve approximately 82%.

5 Mitigation

The above experiments show that our metric-based attack and feature-based attack are effective to some extent, and especially that our attack methods work well on the medical report generation tasks. Meanwhile, to defend the membership inference against multi-modal models, we also evaluate some defense mechanisms to mitigate the threat of membership leakage.

Preventing overfitting. When a target model overfits its training dataset, it may cause different behavior on the training and test datasets, such as higher scores in similarity metrics. An overfitting model can be weaker to the membership inference attack. The widely used techniques to reduce the impact of overfitting in machine learning are data augmentation and l_2 regularization [76]. From the results shown in Figure 10, we find that these two ways do help reduce the risk of membership leakage. However, the defense by data augmentation is not so efficient in some specific cases, for instance, when the target model is trained with IAPR dataset.

Privacy enhanced training. Differential privacy [77–81] can provide formal membership privacy guarantees for each input in the training dataset of a machine learning model. Many differentially private learning algorithms have been proposed. These algorithms add noise to the training data [82], the objective function [83, 80], or the gradient computed by (stochastic) gradient descent during the learning process [78–80]. To evaluate the attack performance with differential privacy training, we directly adopt the Opacus [84], a library for training PyTorch models with differential privacy stochastic gradient descent, on the target model. The differential privacy is set reasonably strong with $\epsilon = 1.3$, given $\delta = 10^{-5}$. In such a case, Figure 10 shows that both metric-based attack and feature-based attack performances are significantly affected. However, with differential privacy, there can be an obvious degradation in the model performance. For instance, the average ROUGE-L score for FR_target model drops from 0.245 to 0.147.

6 Conclusion

In this work, we take the first step to infer the data-level membership in the multi-modal models and propose two attack methods and evaluate them under different assumptions, including metric-based attack and feature-based attack. Extensive experiments show that both our attacks outperform random guesses and the feature-based attack in general outperforms the metric-based attack. Furthermore, we adopt our attack on the pretrained medical report generation model and results show the privacy of patients is also vulnerable. Finally, we evaluate the effectiveness of data augmentation and differential private training mechanisms to defend against our attacks.

The future work can include extending our method to other domains like multi-modality such as voices or videos, extending our methods to other multi-modal tasks such as VQA, and exploring other privacy risks of the multi-modal models such as privacy leakage from the cross-modality generation models.

References

[1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 1

[2] Annamaria Mesaros, Toni Heittola, Tuomas Virtanen, and Mark D Plumbley. Sound event detection: A tutorial. *IEEE Signal Processing Magazine*, 38(5):67–83, 2021. 1

[3] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and Tell: A Neural Image Caption Generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1, 2, 5, 6, 15

[4] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[5] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image Captioning With Semantic Attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[6] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[7] Jyoti Aneja, Aditya Deshpande, and Alexander G. Schwing. Convolutional Image Captioning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5561–5570, 2018.

[8] Marco Pedersoli, Thomas Lucas, Cordelia Schmid, and Jakob Verbeek. Areas of Attention for Image Captioning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[9] Xinpeng Chen, Lin Ma, Wenhao Jiang, Jian Yao, and Wei Liu. Regularizing RNNs for Caption Generation by Reconstructing the Past With the Present. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[10] Quanzeng You, Zhengyou Zhang, and Jiebo Luo. End-to-End Convolutional Semantic Embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[11] Tseng-Hung Chen, Yuan-Hong Liao, Ching-Yao Chuang, Wan-Ting Hsu, Jianlong Fu, and Min Sun. Show, Adapt and Tell: Adversarial Training of Cross-Domain Image Captioner. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[12] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France, 07–09 Jul 2015. PMLR.

[13] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting Image Captioning With Attributes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1, 2

[14] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE S & P*, 2017. 1, 2

[15] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting Training Data from Large Language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association, August 2021. 1, 2

[16] Hongbin Liu, Jinyuan Jia, Wenjie Qu, and Neil Zhenqiang Gong. EncoderMI: Membership Inference against Pre-trained Encoders in Contrastive Learning. *CCS*, 2021. 1, 2

[17] Christopher A. Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. Label-only Membership Inference Attacks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1964–1974. PMLR, 18–24 Jul 2021.

[18] Bo Hui, Yuchen Yang, Haolin Yuan, Philippe Burlina, Neil Zhenqiang Gong, and Yinzhi Cao. Practical Blind Membership Inference Attack via Differential Comparisons. In *NDSS*, 2021. 2

[19] Inken Hagestedt, Yang Zhang, Mathias Humbert, Pascal Berrang, Haixu Tang, Xiaofeng Wang, and Michael Backes. Mbeacon: Privacy-Preserving Beacons for DNA Methylation Data. In *Proceedings of the 26th Annual Network and Distributed System Security Symposium (NDSS)*, 2019.

[20] 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, 2018. 2

[21] Inken Hagestedt, Mathias Humbert, Pascal Berrang, Irina Lehmann, Roland Eils, Michael Backes, and Yang Zhang. Membership Inference Against DNA Methylation Databases. In *2020 IEEE European Symposium on Security and Privacy (EuroS P)*, pages 509–520, 2020.

[22] Xinlei He and Yang Zhang. Quantifying and Mitigating Privacy Risks of Contrastive Learning. *ACM CCS*, 2021.

[23] Jiacheng Li, Ninghui Li, and Bruno Ribeiro. Membership Inference Attacks and Defenses in Classification Models. *Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy*, Apr 2021.

[24] Zheng Li and Yang Zhang. Membership Leakage in Label-Only Exposures. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, CCS '21, page 880–895, New York, NY, USA, 2021. Association for Computing Machinery.

[25] 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. *2019 IEEE Symposium on Security and Privacy (SP)*, May 2019. 2

[26] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Herve Jegou. White-box vs Black-box: Bayes Optimal Strategies for Membership Inference. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5558–5567. PMLR, 09–15 Jun 2019.

[27] Liwei Song and Prateek Mittal. Systematic Evaluation of Privacy Risks of Machine Learning Models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2615–2632. USENIX Association, August 2021.

[28] Liwei Song, Reza Shokri, and Prateek Mittal. Privacy Risks of Securing Machine Learning Models against Adversarial Examples. *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, Nov 2019.

[29] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *CSF*, 2018. 1

[30] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, and Xuyun Zhang. Membership Inference Attacks on Machine Learning: A Survey, 2021. 2

[31] Reza Shokri, Martin Strobel, and Yair Zick. On the Privacy Risks of Model Explanations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 231–241, New York, NY, USA, 2021. Association for Computing Machinery. 2

[32] Yunhui Long, Lei Wang, Diyue Bu, Vincent Bindschaedler, Xiaofeng Wang, Haixu Tang, Carl A. Gunter, and Kai Chen. A Pragmatic Approach to Membership Inferences on Machine Learning Models. In *2020 IEEE European Symposium on Security and Privacy (EuroS P)*, pages 521–534, 2020.

[33] Gaoyang Liu, Chen Wang, Kai Peng, Haojun Huang, Yutong Li, and Wenqing Cheng. SocInf: Membership Inference Attacks on Social Media Health Data With Machine Learning. *IEEE Transactions on Computational Social Systems*, 6(5):907–921, 2019.

[34] Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, and Wenqi Wei. Towards Demystifying Membership Inference Attacks. *arXiv e-prints*, page arXiv:1807.09173, June 2018.

[35] Hanxiao Chen, Hongwei Li, Guishan Dong, Meng Hao, Guowen Xu, Xiaoming Huang, and Zhe Liu. Practical Membership Inference Attack Against Collaborative Inference in Industrial IoT. *IEEE Transactions on Industrial Informatics*, 18(1):477–487, 2022.

[36] Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang. When machine unlearning jeopardizes privacy. *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, Nov 2021.

[37] Congzheng Song and Vitaly Shmatikov. Auditing data provenance in text-generation models. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 196–206, 2019.

[38] Kehao Wang, Zhixin Hu, Qingsong Ai, Quan Liu, Mozi Chen, Kezhong Liu, and Yirui Cong. Membership Inference Attack with Multi-Grade Service Models in Edge Intelligence. *IEEE Network*, 35(1):184–189, 2021. 2

[39] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. Logan: Membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies*, 2019(1):133–152, 2019. 2

[40] Benjamin Hilprecht, Martin Härterich, and Daniel Bernau. Reconstruction and Membership Inference Attacks against Generative Models. *arXiv e-prints*, page arXiv:1906.03006, June 2019.

[41] Kin Sum Liu, Chaowei Xiao, Bo Li, and Jie Gao. Performing Co-membership Attacks Against Deep Generative Models. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 459–467, 2019.

[42] Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. GAN-Leaks: A Taxonomy of Membership Inference Attacks against Generative Models. *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, Oct 2020.

[43] Bingzhe Wu, Shiwan Zhao, ChaoChao Chen, Haoyang Xu, Li Wang, Xiaolu Zhang, Guangyu Sun, and Jun Zhou. Generalization in generative adversarial networks: a novel perspective from privacy protection. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 307–317, 2019.

[44] Sumit Mukherjee, Yixi Xu, Anusua Trivedi, and Juan Lavista Ferres. privGAN: Protecting GANs from membership inference attacks at low cost. *arXiv e-prints*, page arXiv:2001.00071, December 2019.

[45] Ryan Webster, Julien Rabin, Loic Simon, and Frederic Jurie. This person (probably) exists. identity membership attacks against gan generated faces, 2021. 2

[46] Congzheng Song and Ananth Raghunathan. Information leakage in embedding models. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pages 377–390, 2020. 2

[47] Saeed Mahloujifar, Huseyin A. Inan, Melissa Chase, Esha Ghosh, and Marcello Hasegawa. Membership Inference on Word Embedding and Beyond, 2021.

[48] Aleena Thomas, David Ifeoluwa Adelani, Ali Davody, Aditya Mogadala, and Dietrich Klakow. Investigating the Impact of Pre-trained Word Embeddings on Memorization in Neural Networks. In *23rd International Conference on Text, Speech and Dialogue*, brno, Czech Republic, September 2020. 2

[49] Hongkyu Lee, Jeehyeong Kim, Seyoung Ahn, Rasheed Hussain, Sunghyun Cho, and Junggab Son. Digestive neural networks: A novel defense strategy against inference attacks in federated learning. *Computers & Security*, 109:102378, 2021. 2

[50] Jingwen Zhang, Jiale Zhang, Junjun Chen, and Shui Yu. Gan enhanced membership inference: A Passive Local Attack in Federated Learning. *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, pages 1–6, 2020.

[51] Jiale Chen, Jiale Zhang, Yanchao Zhao, Hao Han, Kun Zhu, and Bing Chen. Beyond Model-level Membership Privacy Leakage: an Adversarial Approach in Federated Learning. In *2020 29th International Conference on Computer Communications and Networks (ICCCN)*, pages 1–9, 2020.

[52] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, and Xuyun Zhang. Source inference attacks in federated learning. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 1102–1107. IEEE, 2021.

[53] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 691–706. IEEE, 2019. 2

[54] Minxing Zhang, Zhaochun Ren, Zihan Wang, Pengjie Ren, Zhunmin Chen, Pengfei Hu, and Yang Zhang. Membership Inference Attacks Against Recommender Systems. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 864–879, 2021. 2

[55] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. Leveraging video descriptions to learn video question answering. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 2

[56] Jesus Perez-Martin, Benjamin Bustos, and Jorge Pérez. Improving video captioning with temporal composition of a visual-syntactic embedding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3039–3049, 2021.

[57] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[58] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 2

[59] Zhongyi Han, Benzheng Wei, Stephanie Leung, Jonathan Chung, and S. Li. Towards Automatic Report Generation in Spine Radiology Using Weakly Supervised Framework. In *MICCAI*, 2018. 2

[60] Baoyu Jing, Pengtao Xie, and Eric Xing. On the Automatic Generation of Medical Imaging Reports. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018. 2

[61] Pavel Kisilev, Eli Sason, Ella Barkan, and Sharbell Y. Hashoul. Medical Image Description Using Multi-task-loss CNN. In *LABELS/DLMIA@MICCAI*, pages 121–129, 2016. 2

[62] Xinran Liu, H.R. Tizhoosh, and J. Kofman. Generating binary tags for fast medical image retrieval based on convolutional nets and radon transform. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 2872–2878, 2016.

[63] MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A Comprehensive Survey of Deep Learning for Image Captioning. *ACM Comput. Surv.*, 51(6), feb 2019. 2

[64] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. 4

[65] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 4

[66] Qing-Yuan Jiang and Wu-Jun Li. Deep cross-modal hashing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3232–3240, 2017. 4, 5

[67] Long D Nguyen, Dongyun Lin, Zhiping Lin, and Jiuwen Cao. Deep CNNs for microscopic image classification by exploiting transfer learning and feature concatenation. In *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE, 2018. 5

[68] Hussam Qassim, Abhishek Verma, and David Feinzimer. Compressed residual-VGG16 CNN model for big data places image recognition. In *2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 169–175. IEEE, 2018. 5

[69] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 6, 15

[70] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 6

[71] Ameesh Makadia, Vladimir Pavlovic, and Sanjiv Kumar. A New Baseline for Image Annotation. In David Forsyth, Philip Torr, and Andrew Zisserman, editors, *Computer Vision – ECCV 2008*, pages 316–329, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. 6, 15

[72] Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 2008. 7

[73] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership Inference Attacks From First Principles. *arXiv preprint arXiv:2112.03570*, 2021. 8

[74] Shahbaz Rezaei and Xin Liu. On the Difficulty of Membership Inference Attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7892–7900, 2021. 8

[75] Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Generating Radiology Reports via Memory-driven Transformer. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, August 2021. 8, 15

[76] Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, and Wenqi Wei. Towards Demystifying Membership Inference Attacks. *CoRR abs/1807.09173*, 2018. 9

[77] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, 2006. 9

[78] Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *CCS*, 2015. 9

[79] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *CCS*, 2016.

[80] Bargav Jayaraman and David Evans. Evaluating differentially private machine learning in practice. In *USENIX Security Symposium*, 2019. 9

[81] Milad Nasr, Shuang Song, Abhradeep Thakurta, Nicolas Papernot, and Nicholas Carlini. Adversary Instantiation: Lower Bounds for Differentially Private Machine Learning. In *IEEE S & P*, 2021. 9

[82] John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *FOCS*, 2013. 9

[83] Roger Iyengar, Joseph P Near, Dawn Song, Om Thakkar, Abhradeep Thakurta, and Lun Wang. Towards practical differentially private convex optimization. In *IEEE S & P*, 2019. 9

[84] Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. Opacus: User-friendly differential privacy library in PyTorch. *arXiv preprint arXiv:2109.12298*, 2021. 9, 15

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We discussed the consequences of multi-modal model membership inference in the Introduction.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A] We did not include theoretical results in this research.
   (b) Did you include complete proofs of all theoretical results? [N/A] We did not include theoretical results in this research.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The source code of M^4I attacks is publicly available at `https://github.com/MultimodalMI/Multimodal-membership-inference`. The URL is provided at the end of Abstract.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] The results of multiple times of training will not affect the trend of the results, given limited computing resources.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We provide the computation resource used in our experiment in Supplementary Materials.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We cited all the existing assets used in this research, including code [84, 75], data [69–71], and models [3].
   (b) Did you mention the license of the assets? [Yes] The license of the assets is provided in the Supplementary Materials.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] The URL of M^4I source code is provided at the end of Abstract.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] The three public datasets we used do not contain any personally identifiable information or offensive content.

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] We did not use crowdsourcing or conduct research with human subjects.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] We did not use crowdsourcing or conduct research with human subjects.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] We did not use crowdsourcing or conduct research with human subjects.