Membership inference attack on Multimodal model with partial knowledge

Chidalla Veda Varshith Sai
Kiran Kumar Nimmakayala
Department of Computer Science, Emory University, Atlanta, GA USA

E-mail: cvedava@emory.edu
knimmak@emory.edu

Abstract.

This project investigates Membership Inference Attacks (MIAs) on multimodal models under a partial knowledge setting, where adversaries have access to only image data. Unlike traditional MIA research that assumes access to multiple modalities (e.g., images and text), this work addresses a more realistic adversarial scenario. The goal is to evaluate the privacy vulnerabilities of multimodal models and explore the efficacy of two attack strategies: Confidence-Based Attacks and Top-K Diversity-Based Attacks.

The Confidence-Based Attack exploits the observation that models often generate captions with higher confidence for images seen during training compared to unseen images. By using the model's prediction confidence as a feature, a binary classifier (e.g., Random Forest) distinguishes member images from non-members. The Top-K Diversity-Based Attack, on the other hand, evaluates the variability in the top-k generated captions for an image. Images from the training set typically produce more consistent captions, and this consistency is measured using entropy-based metrics. These features are then used to train attack classifiers such as XGBoost and MLP.

The attack strategies were tested on the Flickr8k and COCO datasets using ResNet-LSTM-based target models. To simulate a realistic attack scenario, shadow models were trained on disjoint subsets of the datasets to generate features for training attack classifiers. The experimental results reveal that the Top-K Diversity-Based Attack, combined with an XGBoost classifier, achieved the best performance with an AUC-ROC of 0.6565 on the COCO dataset, demonstrating its ability to distinguish member images from non-members. The Confidence-Based Attack, while less effective, showed moderate success, with Random Forest achieving an AUC-ROC of 0.5917 on COCO. These findings highlight that membership inference is feasible even under partial knowledge constraints.

The project underscores the privacy risks posed by multimodal models and the need for stronger defenses. It emphasizes that existing defense mechanisms, such as differential privacy and adversarial training, should be revisited to address the vulnerabilities unique to multimodal systems. Furthermore, the results suggest that hybrid attack strategies, which combine confidence-based and diversity-based features, may lead to even greater attack effectiveness. This research provides a foundation for future studies on privacy-preserving multimodal models and highlights the importance of designing more secure AI systems as their use becomes more prevalent in sensitive applications like healthcare, surveillance, and personalized recommendation systems.

1. Introduction

The rise of multimodal models has transformed tasks like image captioning, cross-modal retrieval, and text-to-image generation. By combining data from multiple modalities, such as images and text, these models achieve impressive performance and versatility. However, this advancement comes with a cost — privacy risks. One of the most pressing concerns is Membership Inference Attacks (MIAs), where an adversary attempts to determine if a specific data point was part of a model's training set. This poses serious privacy threats, particularly when dealing with sensitive information like personal photos, medical records, or proprietary datasets.

While MIAs have been well-studied in traditional machine learning models, such as classifiers and GANs, the unique complexities of multimodal models have received less attention. These models often rely on large-scale, publicly available datasets and process multiple types of inputs, making them particularly vulnerable to privacy attacks. Previous research on MIAs for multimodal models assumes that adversaries have access to both image and text data. However, in real-world applications, attackers may have access to only one modality, often the image. This limited access scenario, known as the partial knowledge setting, presents new challenges and reflects the conditions present in services like machine learning as a service (MLaaS), where users typically interact with the model via image inputs alone.

This project aims to address this gap by investigating MIAs on multimodal models under partial knowledge constraints. We evaluate two novel attack strategies: Confidence-Based Attacks and Top-k Caption Diversity Attacks. The Confidence-Based Attack takes advantage of the model's tendency to assign higher confidence to captions generated for images it was trained on, compared to those it has never seen. The Top-k Caption Diversity Attack assesses the variability among the top-k generated captions for an image, operating on the principle that images from the training set tend to produce more consistent captions. By focusing exclusively on image data, these strategies align with the practical limitations faced by adversaries in partial knowledge scenarios.

To evaluate the effectiveness of these attacks, we use the Flickr8k and MS COCO datasets — two widely-used benchmarks in image captioning research. Our target model is a ResNet + LSTM architecture, commonly employed for image captioning tasks. To simulate a realistic attack scenario, we train shadow models on separate subsets of the data, extract features like confidence scores and caption diversity, and train classifiers to distinguish between member and non-member data points. We assess the success of these attacks using key performance metrics, including accuracy, precision, recall, F1 score, and AUC-ROC.

This project aims to shed light on the privacy vulnerabilities of multimodal models in real-world conditions. By focusing on partial knowledge scenarios, we highlight the practical risks posed by these attacks and offer valuable insights into the design of stronger privacy defenses. Our findings emphasize the need for better safeguards in multimodal models, especially as they become more prevalent in sensitive applications

like healthcare diagnostics, recommendation systems, and large-scale AI platforms.

2. Background

Multimodal Models and Privacy Risks Multimodal models, such as those used in image captioning systems, rely on large, publicly available datasets to learn complex relationships between images and text. While this approach achieves state-of-the-art performance, it also raises privacy concerns. A significant risk is that models may memorize portions of their training data, potentially exposing sensitive information during inference. For example, if a model has seen a specific image multiple times during training, it may generate highly specific captions for it, making it easier for an adversary to infer that the image was part of the training set.

These risks are more pronounced in multimodal models, which process multiple types of data (images, text, etc.) together. Each modality introduces its own privacy vulnerabilities. For image-captioning models, the captions themselves can leak information about the training data. If the model produces captions with lower diversity or higher confidence for training images, adversaries can use these patterns to infer membership. This highlights the need to study Membership Inference Attacks (MIAs) in the context of multimodal models.

2.1. Related Work

One of the most influential works in this field is the M4I (Multi-modal Models Membership Inference) framework, which proposed attack strategies that use both image and text data to infer membership. M4I showed that multimodal models are vulnerable to MIAs, even with limited access to the underlying model. However, M4I assumes that adversaries have access to both image and text data, which is not always realistic in real-world applications. This project builds on the M4I approach but narrows the focus to a partial knowledge setting, where only image data is accessible. This approach is more applicable in scenarios like MLaaS platforms, where users typically provide only image inputs.

Additionally, existing Kaggle-based methods for image captioning on the Flickr8k dataset have inspired parts of this project. Prior Kaggle methods used both images and captions for evaluation, but this project modifies the approach to rely solely on image data for MIAs. By doing so, it addresses the unique challenges posed by the partial knowledge setting and proposes new attack strategies tailored to this more realistic threat model.

2.2. Datasets and Models

This project evaluates attack strategies using two widely-used datasets: Flickr8k and MS COCO.

Flickr8k consists of 8,000 images, each paired with five human-generated captions, allowing for quick experimentation and testing. MS COCO contains over 330,000 images, each with 5 to 10 captions, making it a more comprehensive benchmark for large-scale, real-world scenarios. The ResNet + LSTM architecture serves as the target model, where ResNet extracts image features and LSTM generates captions. A shadow model is also trained on a disjoint subset of the data to simulate an adversary's perspective. This shadow model generates confidence scores and diversity metrics for both member and non-member images, which are then used to train classifiers that predict membership status. This setup mirrors real-world adversarial conditions, providing valuable insights into the privacy vulnerabilities of multimodal models.

3. Methods

3.1. Dataset Preparation

The datasets used for training the target models are COCO and Flickr8k. The COCO dataset is split into training and validation sets. The training data is used to train the target model, while the validation set is used to evaluate the attack models. The data is divided into a target training set, which is used to train the target model, a shadow training set that mimics the behavior of the target model and is used for training the attack model, and a query pool containing images that are not part of the target training set but are used to generate query captions for the attack model. Images from COCO and Flickr8k datasets are used to generate captions, and these captions are tokenized and embedded to form the training inputs for the caption generation model. Data augmentation techniques like image resizing and normalization are applied before feeding images into the encoder model.

3.2. Target Model Design

The target model is a caption generation model composed of two main components: an encoder and a decoder. The encoder is a ResNet-152 used to extract image features, and the decoder is a Recurrent Neural Network (RNN) with LSTM cells used to generate captions conditioned on the image features. The target model is trained on the training subset of the COCO dataset, and the model generates captions for each image, which are then used to train the attack models.

3.3. Membership Inference Attack (MIA) Design

We consider two different attack strategies. The first strategy is the confidence-based attack, where for each image, the target model generates top-K captions. The attack extracts the model's prediction probabilities (confidence) for the generated captions. These confidence values are used as features to train a binary classifier (MLP or XGBoost) to distinguish between members and non-members. The second strategy

is the top-K attack, where the diversity of the generated captions is measured. For each image, the entropy of the token predictions at each time step is computed, and the entropy is used as the diversity metric. A higher entropy implies more diversity, while lower entropy indicates that the model tends to generate the same captions repeatedly for member images.

3.4. Attack Model Design

The attack models are binary classifiers designed to predict whether an image was part of the training set (member) or not (non-member). We employ two types of attack models: a Multi-Layer Perceptron (MLP) attack model and an XGBoost attack model. The MLP attack model takes as input feature vectors derived from the confidence-based and top-K approaches. The MLP architecture consists of an input layer whose size depends on the feature vector, two hidden layers with ReLU activations, and an output layer with a single output neuron with a sigmoid activation to predict probabilities. The XGBoost attack model is trained to classify member and non-member images using both confidence-based and top-K features.

3.5. Training Procedure

The target model is trained using cross-entropy loss, with images input to the ResNet encoder and captions generated using the LSTM decoder. A shadow model is trained on a separate subset of data, and its purpose is to imitate the target model's behavior. This allows us to train the attack models without accessing the target model's training set. For the attack models, input features are extracted from member and non-member images, and the member images are labeled as 1, while non-member images are labeled as 0. The extracted features are used to train MLP and XGBoost attack models. The models are trained using binary cross-entropy loss, and evaluation is performed to measure the attack's success.

3.6. Evaluation Metrics

The effectiveness of the attack models is measured using accuracy, precision, recall, F1-score, and AUC-ROC metrics. Accuracy measures the proportion of correct predictions, while precision measures the proportion of correctly classified members relative to all predicted members. Recall measures how well the attack detects actual members from the dataset, and the F1-score is the harmonic mean of precision and recall. The AUC-ROC curve plots the True Positive Rate (TPR) vs. the False Positive Rate (FPR) to visualize the attack's ability to distinguish members from non-members.

3.7. Experimental Setup

All models are trained on an NVIDIA GPU to ensure faster computations. The deep learning models are implemented using PyTorch, while attack models are implemented

using XGBoost and MLP. Libraries such as NumPy, Matplotlib, and SciPy are used for feature extraction and visualization. The learning rate for the MLP is set to 0.001, with a batch size of 32 and a total of 20 epochs for attack model training.

4. Results

4.1. Flickr8k Dataset

4.1.1. Confidence-Based Attack The performance of the confidence-based attack on the Flickr8k dataset is summarized in Table 1, and the corresponding AUC-ROC curve is shown in Figure 1.

Table 1: Performance Metrics for Confidence-Based Attack on Flickr8k Dataset

Class	**Precision**	**Recall**	**F1-Score**
Non-Member (0)	0.55	0.59	0.57
Member (1)	0.56	0.52	0.54
Accuracy	0.56		
Macro Avg.	0.56	0.56	0.56
Weighted Avg.	0.56	0.56	0.56

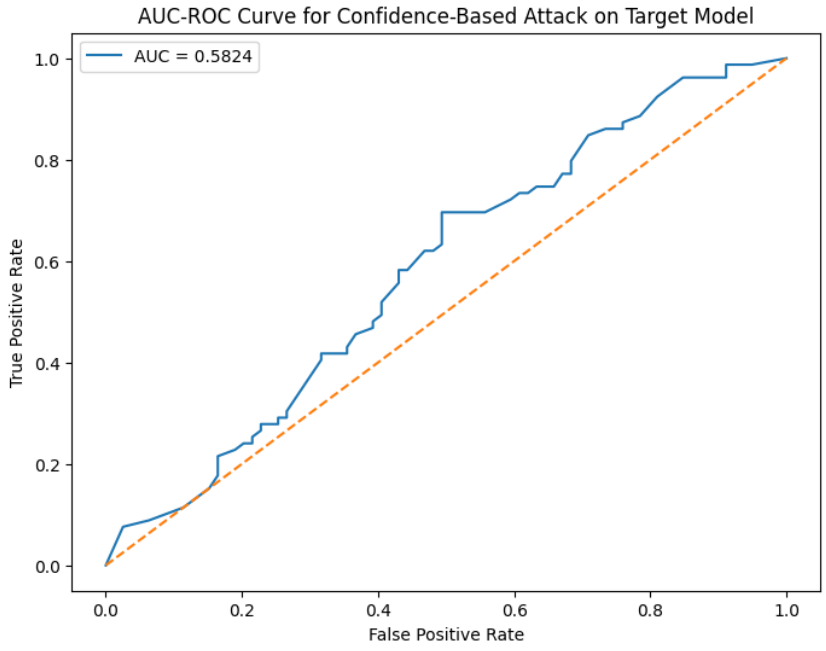

Figure 1: AUC-ROC Curve for Confidence-Based Attack on Flickr8k Dataset

The AUC-ROC for the confidence-based attack was 0.5824 (Figure 1). These results indicate a moderate ability to distinguish members from non-members using confidence-based features.

4.1.2. Top-K Diversity Attack The performance of the top-K diversity attack on the Flickr8k dataset is presented in Table 2, and the AUC-ROC curve is displayed in Figure 2.

Table 2: Performance Metrics for Top-K Diversity Attack on Flickr8k Dataset

Class	**Precision**	**Recall**	**F1-Score**
Non-Member (0)	0.51	0.54	0.53
Member (1)	0.52	0.48	0.50
Accuracy	0.51		
Macro Avg.	0.51	0.51	0.51
Weighted Avg.	0.51	0.51	0.51

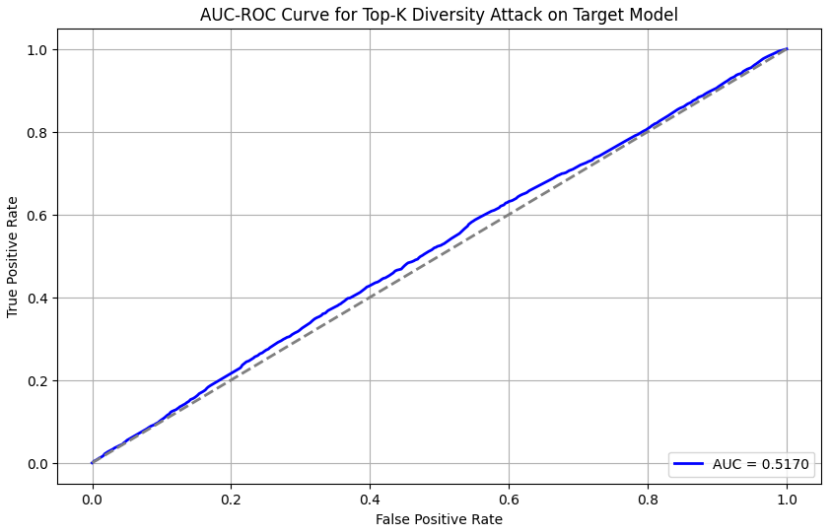

Figure 2: AUC-ROC Curve for Top-K Diversity Attack on Flickr8k Dataset

The top-K diversity attack achieved an AUC-ROC of 0.5170 (Figure 2). While the overall performance is lower compared to the confidence-based attack, the results demonstrate that diversity-based metrics provide additional insights into membership inference.

4.2. COCO Dataset

4.2.1. Confidence-Based Attack : The confidence-based attack on the COCO dataset was evaluated using two models: Logistic Regression and Random Forest. Logistic Regression performed poorly, showing results close to random guessing, while Random Forest demonstrated much better performance. The comparison of both models is summarized in Table 3, and the AUC-ROC curve for the Random Forest model with balanced data is shown in Figure 3.

Table 3: Performance Metrics for Confidence-Based Attack on COCO Dataset

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	0.6715	0.6715	1.0000	0.8035	0.5026
Random Forest (Balanced)	0.5579	0.5375	0.7894	0.6396	0.5917

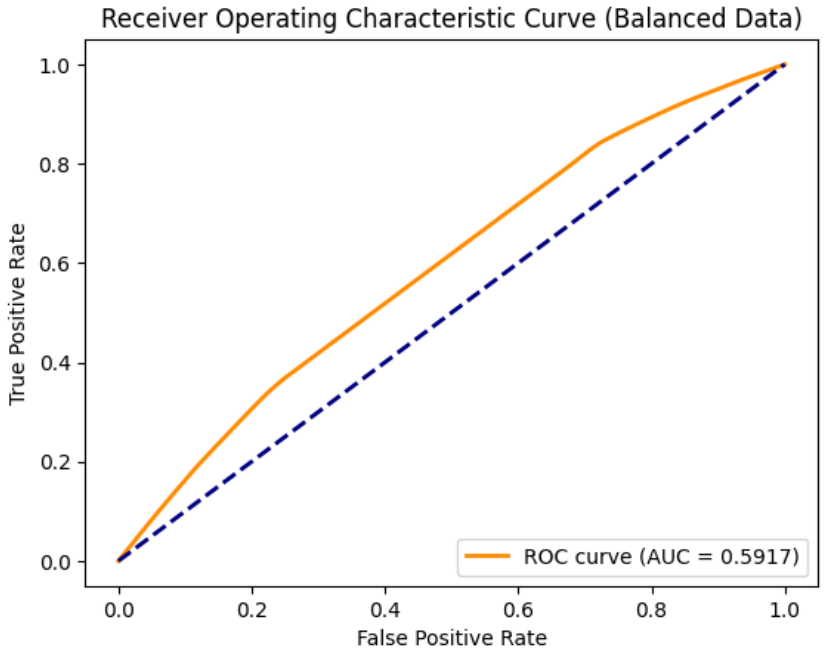

Figure 3: AUC-ROC Curve for Confidence-Based Attack using Random Forest on COCO Dataset (Balanced Data)

As shown in Table 3, Logistic Regression achieved an accuracy of 67.15% but had an AUC-ROC of 0.5026, indicating random guessing behavior. In contrast, Random Forest achieved an F1-score of 0.6396 and an AUC-ROC of 0.5917 (Figure 3), demonstrating better reliability and effectiveness in distinguishing members from non-members.

4.2.2. Top-K Diversity-Based Attack : The top-K diversity-based attack was evaluated on the COCO dataset using three models: Random Forest, Neural Network (NN), and XGBoost. The results indicate that XGBoost outperformed the other models significantly, while Random Forest performed poorly, and the Neural Network exhibited behavior close to random guessing. Table 4 summarizes the results, and the AUC-ROC curve for the XGBoost model is shown in Figure 4.

The performance metrics in Table 4 demonstrate the superiority of XGBoost, achieving an F1-score of 0.8092 and an AUC-ROC of 0.6565. In contrast, Random Forest struggled with a low AUC-ROC of 0.3241, and the Neural Network behaved similarly to random guessing with an AUC-ROC of 0.5001. The AUC-ROC curve for XGBoost in Figure 4 illustrates its strong ability to distinguish members from non-

Table 4: Performance Metrics for Top-K Diversity-Based Attack on COCO Dataset

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Random Forest	0.4420	0.2290	0.0490	0.0807	0.3241
Neural Network (NN)	0.5000	0.5000	1.0000	0.6667	0.5001
XGBoost	0.6859	0.6842	0.9903	0.8092	0.6565

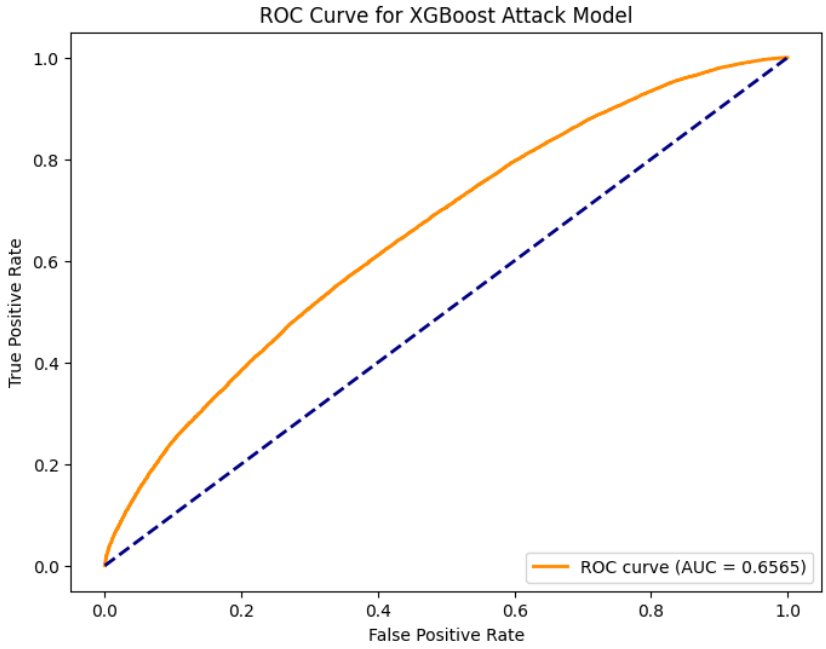

Figure 4: AUC-ROC Curve for Top-K Diversity-Based Attack using XGBoost on COCO Dataset

members, highlighting its effectiveness in the top-K diversity-based attack.

5. Discussion

The results of the membership inference attacks (MIAs) on the COCO and Flickr8k datasets reveal critical insights into the efficacy of various attack strategies and models. The performance of the confidence-based and top-K diversity-based attacks was examined across multiple classifiers, leading to significant observations about the strengths and limitations of each approach.

The confidence-based attack demonstrated moderate success in distinguishing members from non-members. On the COCO dataset, Random Forest outperformed Logistic Regression, achieving an AUC-ROC of 0.5917 compared to Logistic Regression's near-random AUC-ROC of 0.5026. This highlights the importance of selecting an appropriate classifier when leveraging confidence scores. Random Forest's ability to handle complex feature interactions proved beneficial in this context, as it was better suited to capturing subtle differences in confidence distributions between members and

non-members.

While the accuracy and F1-scores for confidence-based attacks were modest, the AUC-ROC curves suggest that these features provide a viable signal for membership inference. However, the overall moderate performance indicates that confidence-based attacks alone may not be sufficient to fully exploit vulnerabilities in the target models, especially for more robust datasets like COCO.

The top-K diversity-based attack provided more nuanced results. Among the three models evaluated on the COCO dataset, XGBoost consistently outperformed Random Forest and Neural Networks, achieving an AUC-ROC of 0.6565 and an F1-score of 0.8092. These results highlight XGBoost's capability to leverage diversity metrics effectively, likely due to its gradient-boosted decision trees' ability to capture non-linear relationships in the data.

Random Forest performed poorly in the top-K diversity-based attack, with an AUC-ROC of only 0.3241, indicating that it struggled to differentiate between members and non-members based on diversity metrics. Similarly, the Neural Network's performance was near-random, with an AUC-ROC of 0.5001, suggesting that the network failed to learn meaningful patterns from the diversity-based features.

The superior performance of XGBoost highlights the potential of diversity-based attacks when paired with robust classifiers. This also underscores the importance of feature engineering and model selection in MIA research, as the choice of attack model has a significant impact on performance.

5.1. Comparative Insights and Limitations

The comparative analysis of confidence-based and top-K diversity-based attacks reveals that both strategies offer unique strengths. Confidence-based attacks are relatively straightforward to implement and can leverage simple metrics such as prediction probabilities. However, their moderate performance suggests that they may not fully capture the intricacies of membership inference, particularly in datasets with high variability.

In contrast, top-K diversity-based attacks take advantage of semantic-level information by assessing the variability in generated outputs, providing a richer set of features for classifiers like XGBoost. However, this approach can be more computationally intensive, as it involves generating and analyzing multiple captions for each image. Additionally, the effectiveness of diversity-based attacks heavily relies on the model's ability to generate noticeably distinct outputs for non-members, which might not always occur.

5.2. Implications for Privacy and Defense Mechanisms

The findings of this study have important implications for privacy in multimodal models. The moderate success of MIAs highlights the potential vulnerabilities in current systems, particularly when models are trained on sensitive data. Defense mechanisms such as

differential privacy and regularization can play a crucial role in mitigating these risks, but their implementation often comes at the cost of reduced model performance.

The results also emphasize the need for robust evaluation frameworks that consider both attack efficacy and the trade-offs associated with defense mechanisms. Future work could explore the integration of adversarial training and advanced privacy-preserving techniques tailored specifically for multimodal models.

5.3. Future Directions

This study highlights several areas for future research. First, exploring hybrid attack strategies that combine confidence-based and diversity-based features may yield better performance by leveraging the strengths of both approaches. Second, examining the impact of varying hyperparameters, such as the value of k in top-K diversity-based attacks, could provide deeper insights into optimizing these methods. Lastly, evaluating the efficacy of these attacks on more diverse datasets and larger multimodal models, such as CLIP and DALL-E, would help generalize the findings and assess their applicability to real-world scenarios.

6. Conclusions

This project investigated membership inference attacks (MIAs) on multimodal models in a partial knowledge scenario, where adversaries have access to only image data. Two key attack strategies, confidence-based and top-K diversity-based attacks, were evaluated on the COCO and Flickr8k datasets using classifiers such as Logistic Regression, Random Forest, Neural Networks, and XGBoost. The findings provide critical insights into the strengths and limitations of these approaches and their implications for privacy in multimodal systems.

In the partial knowledge setting, confidence-based attacks demonstrated moderate success in distinguishing members from non-members. While Logistic Regression performed poorly, Random Forest achieved better results, particularly on the COCO dataset, due to its ability to capture complex feature interactions. Top-K diversity-based attacks showed stronger potential, leveraging semantic-level information to assess variability in generated outputs. XGBoost consistently outperformed other classifiers, achieving the highest AUC-ROC scores and F1-scores. This highlights the importance of selecting robust classifiers and extracting meaningful features in partial knowledge settings.

The results reveal that while both attack strategies exploit vulnerabilities in multimodal models under partial knowledge conditions, their overall effectiveness remains moderate. This underscores the need for further research into more sophisticated attack methodologies and enhanced feature extraction techniques tailored to adversaries with limited access. Additionally, the findings emphasize the importance of implementing robust defense mechanisms, such as differential privacy, adversarial

training, and regularization, to mitigate these risks without significantly compromising model performance.

References

Ko, M., Jin, M., Wang, C. and Jia, R. (2023). Practical membership inference attacks against large-scale multi-modal models: A pilot study.
URL: *https://arxiv.org/abs/2310.00108*

Salem, A., Zhang, Y., Humbert, M., Berrang, P., Fritz, M. and Backes, M. (2019). Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models.
URL: *https://arxiv.org/abs/1806.01246v2*

Shokri, R., Stronati, M., Song, C. and Shmatikov, V. (2017). Membership inference attacks against machine learning models.
URL: *https://arxiv.org/abs/1610.05820*

Wu, Y., Yu, N., Li, Z., Backes, M. and Zhang, Y. (2022). Membership inference attacks against text-to-image generation models, *arXiv preprint arXiv:2210.00968* .
URL: *https://arxiv.org/abs/2210.00968v1*