Final Project Evaluation

Aditi Singh – 23b1053

Sabil Ahmad – 23b1057

4th May 2025

Problem Statement

Input:

Pretrained Classification Model on Image datasets

Output:

Generate images on which the input model was trained

Example:

Input: A classifier Model Trained on

Output: A classifier model trained on CelebA to predict facial attributes.

Brief explanation of the problem tackled in the paper or project.

1. **Better Loss Function**: Existing MI attacks use basic identity losses that fail to leverage full model output information
2. **Overfitting in MI**: Inversion models overfit to the target model, generating unrealistic samples lacking true data semantics

Clearly describe what the model/system is supposed to solve

The system aims to generate realistic approximations of the training data from a pre trained classifier, thereby demonstrating and analyzing the privacy risks posed by MI attacks.

Motivation

- Why did you choose this problem?

1. Model Inversion (MI) attacks pose a serious threat to privacy by reconstructing sensitive training data.
2. Understanding and defending against MI helps in building more secure and trustworthy ML systems.

- What makes it interesting, impactful, or challenging?

1. It's challenging due to the complex interplay between model generalization and memorization.
2. Impactful as it directly relates to data privacy in widely used AI systems like biometrics and medical imaging.

- Is there a real-world application or gap this project addresses?

1. Yes, it addresses the security gap in ML deployments where models can leak training data
2. Real-world applications include safeguarding user data in face recognition systems and patient records in healthcare AI.

Literature Review

Brief background and prior work

1. Model Inversion (MI) attacks aim to reconstruct training data (e.g., faces) from model outputs, raising privacy concerns.
2. The Secret Revealer used GANs to exploit overfitted classifiers and reconstruct high-fidelity images.
3. KEDMI introduced logit loss and distribution matching to improve inversion accuracy and semantic alignment.
4. Re-thinking MI further advanced this by identifying sub-optimal loss functions and MI overfitting, proposing model augmentation and refined objectives.

Key papers or methods that inspired your approach

1. The Secret Revealer (Zhu et al.) — pioneered GAN-based model inversion.
2. KEDMI (Zhang et al.) — introduced knowledge distillation and logit loss for better identity preservation.
3. Re-thinking MI (Li et al.) — highlighted flaws in existing MI techniques and introduced robust improvements via model augmentation and new loss variants.

Dataset

Example Data Instance

A pretrained classification model trained on LFW face images to classify facial attributes (e.g., smiling vs. not smiling).

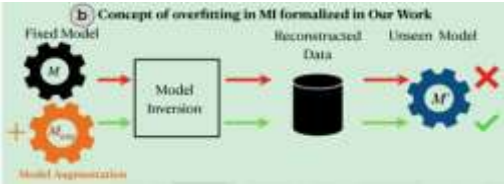
Output

Generated face image resembling training data that the model was trained on, e.g., a smiling young woman with similar features to real LFW samples.

Statistics:

Dataset	Samples	Classes	Image Size
LFW	~13,000	5749 unique identities	128*128
CelebA	~200,000	40 binary attributes	128*128
FFHQ	~70,000	N/A(Used for GAN Training)	128*128

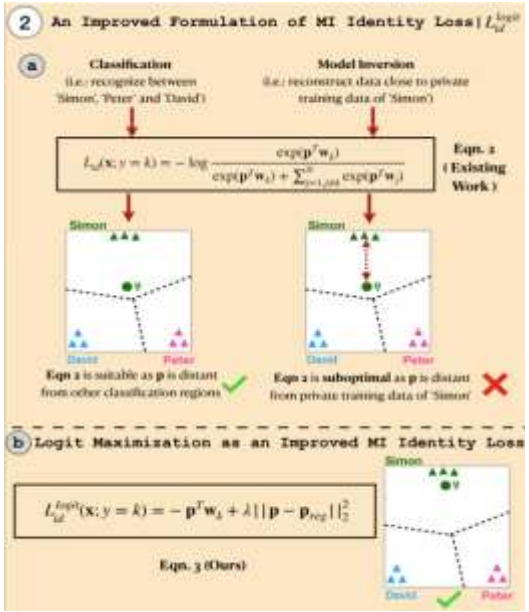
Method/Technique



Model Augmentation

Model augmentation reduces MI overfitting by using **knowledge distillation** to train multiple **augmented models** on public data to mimic the target model. During inversion, identity loss is averaged across the target & augmented models, encouraging reconstructions that generalize better and preserve identity semantics rather than just copying the model's parameters.

$$L_{id}^{aug}(\mathbf{x}; y) = \gamma_t \cdot L_{id}(\mathbf{x}; y, M_t) + \gamma_{aug} \cdot \sum_{i=1}^{N_{aug}} L_{id}(\mathbf{x}; y, M_{aug}^{(i)})$$



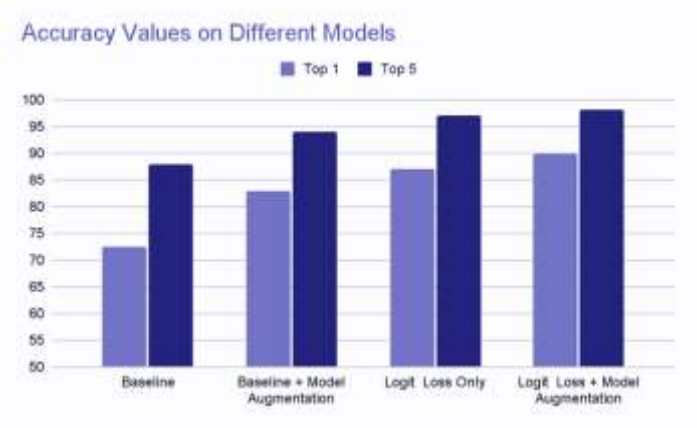
Improved Identity Loss via Logit Maximization

Improved Identity loss since the traditional loss used in model inversion attacks are sub-optimal. To address this, a new identity loss function is used that directly **maximizes the logit corresponding** to the target class, leading to more accurate reconstructions of private training data.

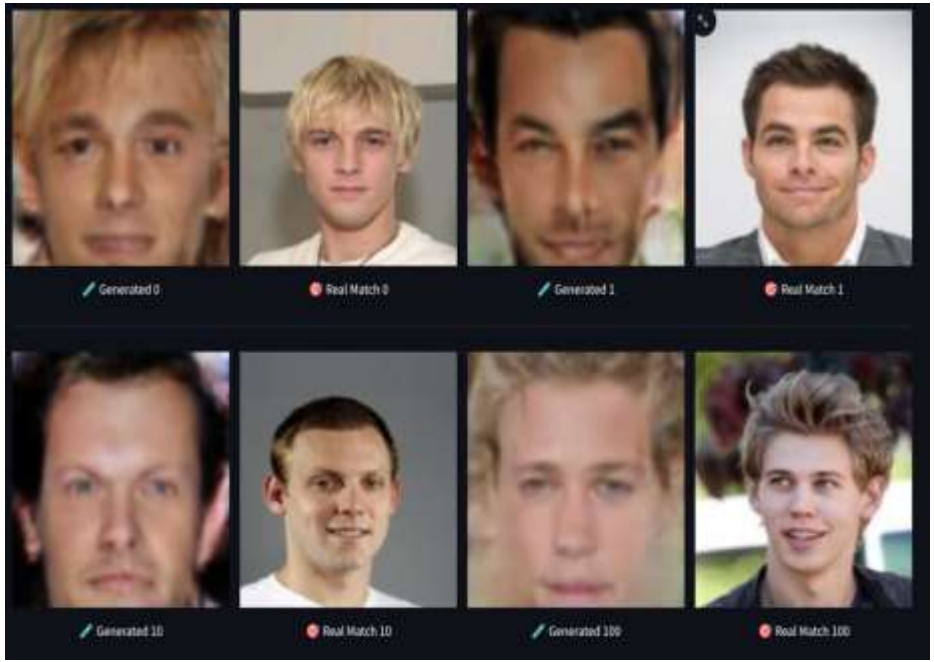
$$L_{id}^{logit}(\mathbf{x}; y = k) = -\mathbf{p}^T \mathbf{w}_k + \lambda ||\mathbf{p} - \mathbf{p}_{reg}||_2^2$$

This loss pushes the features of the generated image to lie in the direction of the **class weight vector** for the desired identity, effectively "pulling" the image representation toward class kkk in feature space.

Results



Images Generated by Our Model vs The Real Match



Analysis

- **Logit loss consistently outperforms cross-entropy** in terms of Top-1 and Top-5 attack accuracy, confirming its effectiveness as a more expressive identity loss for model inversion attacks.
- **Model augmentation improves performance** when used with both loss functions, indicating that it helps reduce overfitting and enhances generalization, especially in MI settings without access to private data.
- The **best trade-off** is seen in the **logit loss + augmentation** variant, which maintains high identity reconstruction accuracy while keeping KNN distance low and FID moderate, suggesting good semantic fidelity and visual quality.
- Interestingly, while **cross-entropy with augmentation** has a better FID (25.173), its lower accuracy suggests that **FID alone doesn't capture identity preservation**, reinforcing the need for metrics like KNN and logit-based loss in MI evaluation.

Error analysis

- Incorrect Outputs with **MSE Loss**: When MSE loss was used (instead of identity-aware losses like cross-entropy or logit loss), the model generated blurry or identity-inconsistent images.

- Reason: **MSE focuses on pixel-wise similarity** rather than semantic identity. As a result, the reconstructions may look smooth or averaged but lack distinct features that match the true class label, causing confusion during inversion.

Improvements over the paper

Expanded Loss Function Evaluation:

Explored novel loss configurations like MSE, Cross Entropy, and **Logit Loss**—both independently and with **Model Augmentation**—to better understand their effects on MI performance.

Visual Evaluation Pipeline & Dataset Changes:

Experimented with **LFW** for training the classifier to ensure better generalization, and added a **neural matching network** for visualization along with metrics like **Top-k Accuracy**, **KNN Distance**, and **FID Score** for robust analysis.

Learnings

1. **Understanding Model Inversion Attacks**:
   Gained deep insights into how adversaries can reconstruct private training data using MI techniques and the significance of identity loss in this process.

2. **Hands-on with GANs & Loss Functions**:
   Learned to leverage GANs for image reconstruction, and how modifying loss functions (e.g., Logit vs Cross Entropy) significantly affects inversion quality.

3. **Impact of Small Design Tweaks:**
   Realized how strategies like Model Augmentation via Knowledge Distillation and thoughtful dataset choices can drastically improve robustness and performance in adversarial setting

Demo

To see the results just run the command on the server

```
streamlit run app.py --server.address 0.0.0.0 --server.port 8510
```

To see the results and metrics run, the results would be loaded in the attack_results/ directory

```
python --customizations evaluation.py
```

Summary and Conclusion

- **Problem Recap :** Tackled the challenge of **Model Inversion (MI) attacks**, where adversaries attempt to reconstruct private training data from ML models. This exposes serious privacy risks, especially in sensitive domains like facial recognition or healthcare.

- **Key Insight :** Found that using **Logit-based Identity Loss** and **Model Augmentation** significantly improves reconstruction accuracy while reducing overfitting. These small architectural and training tweaks outperformed traditional methods like cross-entropy loss alone.

- **Future Work :** Explore **MI defenses**, test on **larger and more diverse datasets**, and develop better **metrics or models for semantic similarity** in reconstructions. Additionally, adapting these techniques to other domains (like NLP) could broaden their impact.