The Many-faced God: Attacking Face Verification System with Embedding and Image Recovery

Mingtian Tan, Zhe Zhou, Zhou Li

Fudan University, China
University of California, Irvine, USA
Background

Face Verification System

- border control
- company entrance
- mobile device
Face Verification System

- border control
- company entrance
- mobile device

Embeddings
Background

Data Privacy Leakage Task

- Membership Inference Attack

- Attribute Inference Attack

Distinguish Training Data and Test Data

Complete Data from part of information
Our Data Privacy Leakage

- Information leakage from FVS

- Embedding recovery from leakage

- Image recovery with embedding
Background

Our Data Privacy Leakage

- Information leakage from FVS
- Embedding recovery from leakage
- Image recovery with embedding
Methodology

Get Information leakage from FVS

self-service FVS in Chinese Entry & Exit Bureau

Attacker - ID number

Victim

Embedding A

Embedding V

Distance
Recover Embedding from leakage

\[ \| \hat{e}_s - e_1 \| = d_1 \]
\[ \| \hat{e}_s - e_2 \| = d_2 \]
\[ \ldots \]
\[ \| \hat{e}_s - e_n \| = d_n \]

Equation Solving

SVD Rank Retained
Methodology

Attack Model Overview

- **Attacker**: ID = “Mary”
  - Query images
  - Attackers embedding model
    - White-box (f)
    - Black-box (f web interface)
    - No-box (f)

- **Enrolled**: [0.91, 0.88, 0.22, 0.35] Distances
  - Solving $\frac{e_s^T}{e_s} \cdot \frac{e_s}{e_s} + A \cdot e_s + D = 0$
  - [0, 0.02, 0, ...]
  - [0.13, 0, 0.2, ...]
  - [0.15, 0.02, 0, ...]
  - [0.66, 0.2, 0, ...]
  - Embeddings

- **Discriminator**
  - Training

- **Loss**
  - $w_r L_r$
  - $w_d L_d$
  - $w_e L_e$

- **Recovered image**
  - White-box (f)
  - Black-box (f)

- **EmbRev**

**embedding-reverse GAN (erGAN)**
Methodology

Generator of erGAN

Extract information from Embedding
Recover the Image from Information
## Evaluation

### Embedding Model (EM)

<table>
<thead>
<tr>
<th>Model</th>
<th>Emb. Dim.</th>
<th>Distance Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residual Inception Network</td>
<td>1792</td>
<td>Cosine</td>
</tr>
<tr>
<td>Clarifai Online Face Embedding [10]</td>
<td>1024</td>
<td>Cosine</td>
</tr>
<tr>
<td>Facenet 20180402-114759 [51]</td>
<td>512</td>
<td>Cosine</td>
</tr>
<tr>
<td>Facenet 20170512-110547 [51]</td>
<td>128</td>
<td>L2</td>
</tr>
</tbody>
</table>

### Embedding Recovery Evaluation

- **LFW Dataset**

### Face Recovery Evaluation

- **celebA dataset**
Embedding Recovery Evaluation

EM : Facenet 128 model
Face Recovery Evaluation

Original

Facenet-128

Facenet-512

Clarifai-1024

WebRes-1024

White-box

Le
Evaluation

Face Recovery Evaluation

<table>
<thead>
<tr>
<th>Model</th>
<th>Blackbox Baseline</th>
<th>White-box</th>
<th>Blackbox $L_e$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acc.</td>
<td>93.07% 97.23% 98.63% 93.87%</td>
<td>94.20%</td>
<td>96.23%</td>
</tr>
<tr>
<td>FID</td>
<td>114.11 157.47 33.94 49.39</td>
<td>86.00</td>
<td>61.25</td>
</tr>
</tbody>
</table>

The second row show the acceptance rate of the images recovered
The third shows the FID of the generated images (smaller is better).
Evaluation

Face Recovery Evaluation

Stability

![Graph showing probability density vs distance with different labels: Recovered, Diff, Same.](image)
Evaluation

Face Recovery Evaluation

Images Recovered with different number of queries.
Thank you !