Model inversion attacks on facial and speaker recognition models

Authors:
Alexandre Duhamel (afd2153)
Edward Lucyszyn (el3331)
Yehiya Monla (yhm2102)

Professor:
Gary Kazantsev

May 6, 2025

Industrial Engineering and Operations Research
Columbia | Engineering

Topic

Can we "hack" a facial recognition model and retrieve its original training data?

Facial recognition model in Mission Impossible [6]

Original paper and work

- Our choice was: *"Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures"* (Fredrikson *et al.* [2]).
- Very first paper to introduce concept of "model inversion attacks"!

Example of model inversion attack in a facial recognition model [2]

Implementation

- Started from an unofficial `GitHub` repository [7]
- Face dataset: AT&T ("ORL") [1]

We inverted three facial-recognition models: Softmax, MLP and DAE (denoising autoencoder).

Model	**Paper's error**	**Our model's error**
Softmax	7.5%	$8.1 \pm 1.2\%$
MLP	4.2%	$4.5 \pm 0.4\%$
DAE	3.3%	$8.6 \pm 1.0\%$

Each was attacked in two threat settings:

- **White–box:** Full access to the model's weights and gradients. We perform direct gradient-based inversion (momentum-SGD) on the loss. (Possible if the model is on your phone for example).
- **Black–box:** Only query access to output confidence scores.

How to invert a model? (white-box setting)

Algorithm 1 Inversion attack for facial recognition models.

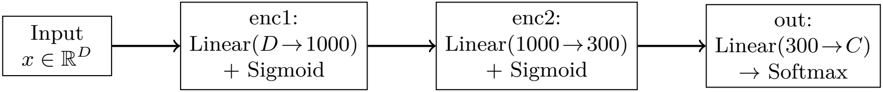
function MI-FACE($label, \alpha, \beta, \gamma, \lambda$)
  $c(\mathbf{x}) \stackrel{\text{def}}{=} 1 - \tilde{f}_{label}(\mathbf{x}) + \text{AUXTERM}(\mathbf{x})$
  $\mathbf{x}_0 \leftarrow \mathbf{0}$
  for $i \leftarrow 1 \ldots \alpha$ **do**
    $\mathbf{x}_i \leftarrow \text{PROCESS}(\mathbf{x}_{i-1} - \lambda \cdot \nabla c(\mathbf{x}_{i-1}))$
    if $c(\mathbf{x}_i) \geq \max(c(\mathbf{x}_{i-1}), \ldots, c(\mathbf{x}_{i-\beta}))$ **then**
      break
    if $c(\mathbf{x}_i) \leq \gamma$ **then**
      break
  return $[\arg\min_{\mathbf{x}_i}(c(\mathbf{x}_i)), \min_{\mathbf{x}_i}(c(\mathbf{x}_i))]$

Inversion attack algorithm for facial recognition models [2]

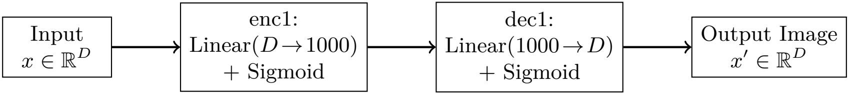
- Gradient based search approach;
- AuxTerm($\mathbf{x}$) is an extra regularizer or projection penalty to keep the image realistic;
- Stop if we reach α iterations, if the loss hasn't improved over the last β iterations or if the loss itself drops below a threshold γ.

How to invert a model? (white-box setting for DAE)

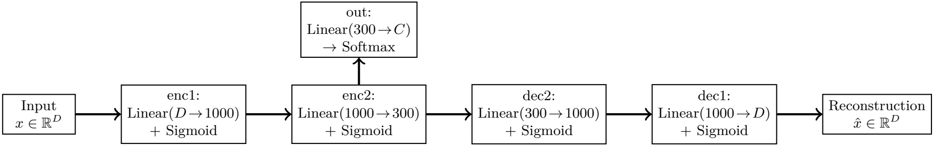
Facial recognition model for DAE:



Then, two things for two steps for the mode inversion attack:



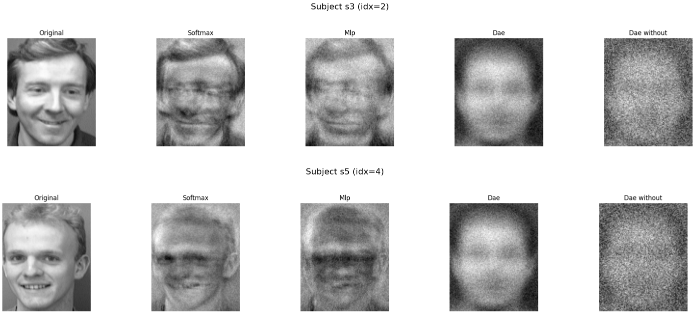
and then, for the inversion:



Results (white-box setting)

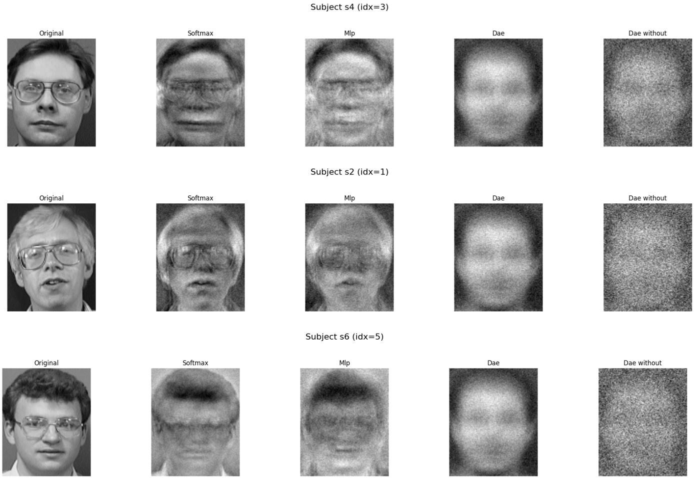
Initialization:

- **Softmax:** $\alpha = 50\,000, \beta = 1000, \gamma = 10^{-4}, \lambda = 0.05, \mu = 0.95$
- **MLP:** same as Softmax
- **DAE:** $\lambda = 0.1, \mu = 0.9, \alpha = 5\,000, \beta = 100, \gamma = 10^{-3}$



Examples of inversion attacks we have generated

Results (white-box setting)



Examples of inversion attacks we have generated

How to invert a model? (black-box setting)

Algorithm 1 Inversion attack for facial recognition models.

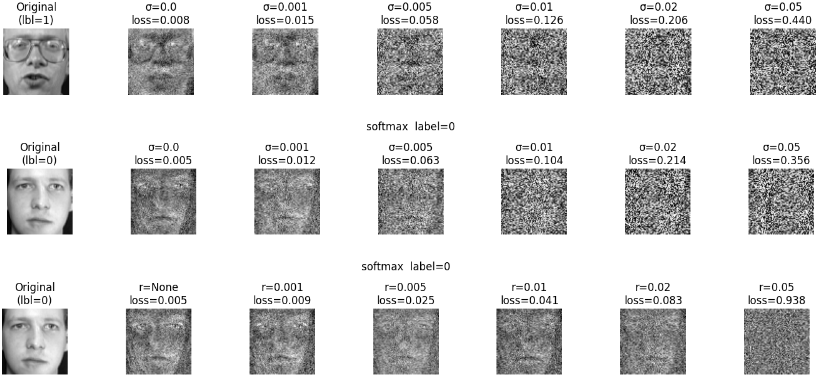
function MI-FACE($label, \alpha, \beta, \gamma, \lambda$)
    $c(\mathbf{x}) \stackrel{\text{def}}{=} 1 - \tilde{f}_{label}(\mathbf{x}) + \text{AUXTERM}(\mathbf{x})$
    $\mathbf{x}_0 \leftarrow \mathbf{0}$
    for $i \leftarrow 1 \ldots \alpha$ **do**
        $\mathbf{x}_i \leftarrow \text{PROCESS}(\mathbf{x}_{i-1} - \lambda \cdot \nabla c(\mathbf{x}_{i-1}))$
        if $c(\mathbf{x}_i) \geq \max(c(\mathbf{x}_{i-1}), \ldots, c(\mathbf{x}_{i-\beta}))$ **then**
            break
        if $c(\mathbf{x}_i) \leq \gamma$ **then**
            break
    return $[\arg\min_{\mathbf{x}_i}(c(\mathbf{x}_i)), \min_{\mathbf{x}_i}(c(\mathbf{x}_i))]$

Inversion attack algorithm for facial recognition models [2]

- Instead of computing the exact gradient $\Longrightarrow$ approximate it!
- For small $\varepsilon > 0$, we have:

$$\frac{\partial c}{\partial y}(x) \simeq \frac{c(x + \varepsilon\, y) \;-\; c(x - \varepsilon\, y)}{2\,\varepsilon\,||y||}.$$

Results (black-box setting)



Examples of inversion attacks we have generated

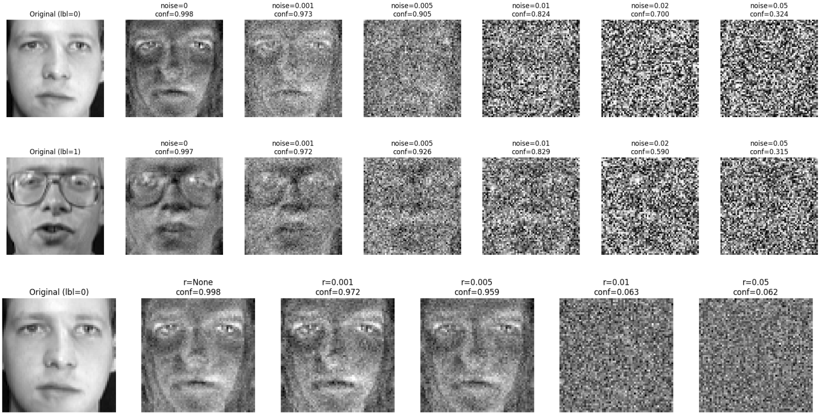
Countermeasures

How to prevent the model from model inversion attacks?

- **Rounding confidences**
  - Quantize
    $$p_j \mapsto \left\lfloor \frac{p_j}{r} \right\rfloor \times r \quad \text{then re-normalize.}$$
  - Blocks SPSA inversion for $r \geq 10^{-2}$, with $< 0.5\%$ accuracy drop.
- **Gaussian output noise**
  - Add $\epsilon \sim \mathcal{N}(0, \sigma^2)$ to each p_j, clip to $[0, 1]$, re-normalize.
  - Faces unrecognizable for $\sigma \geq 0.01$, with $\approx 1\%$ accuracy drop.

Defense	**Block threshold**	**Accuracy drop**
None	–	0%
Rounding (r)	$r \geq 10^{-2}$	$< 0.5\%$
Gaussian noise	$\sigma \geq 0.01$	$\approx 1\%$

Examples of countermeasures (black-box setting)



Examples of inversion countermeasures we have generated

Speaker Recognition Model

- Can we perform the same type of model inversion on a speaker recognition model (i.e., starting from audio training data)?
- Can we create voice deepfakes of individuals in the training data using the inverted audio samples?

Speaker Recognition Model

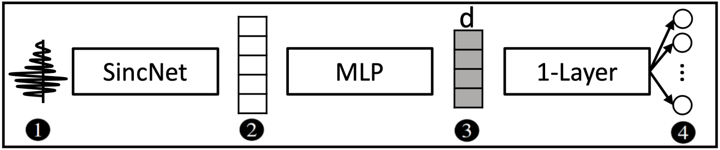
An article already addressed all of our questions: *"Introducing Model Inversion Attacks on Automatic Speaker Recognition"* [4].

- **Inverting audio samples:** they achieved 90.48% accuracy with inverted audio samples reconstructed via model inversion attacks starting from Laplace noise.
- **Creating deepfakes:** the generated audio samples are not perceptually close to the originals for human listeners, but they are close enough to fool automated detection systems. However, using a vocoder, the authors were able to generate a few high-quality spoofed audio samples that resembled the original speaker.

This was done entirely in a **white-box setting** — could it also be **feasible in a black-box setting**?

Speaker recognition model: implementation

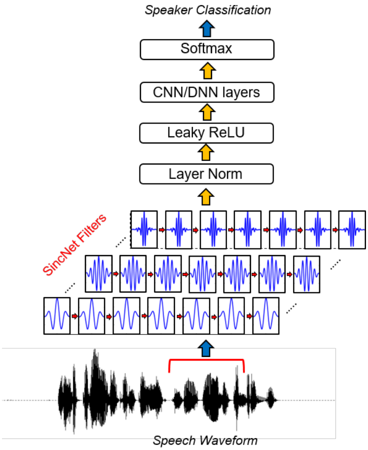
- To try model inversion in a black-box setting we had to create a speaker recognition model!
- **Dataset:** TIMIT [3] $\rightarrow$ broadband recordings of 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences.



Speaker recognition model described in [4]

Speaker recognition model: implementation

- **SincNet:** neural architecture for processing raw audio samples [5].



SincNet architecture [5]

Model Inversion Attack for Speech Recognition

Algorithm 1 Invert SincNet via Gradient Descent

Require: Pre-trained speaker-ID model f, target label c, iterations N, learning rate η
Ensure: Generated waveform x_N

Initialize: $x \leftarrow \mathcal{N}(0, I)$ ▷ Gaussian noise
for $t = 0 \rightarrow N - 1$ **do**
    $z \leftarrow f(x)$ ▷ logits from model
    $p \leftarrow \mathrm{softmax}(z)$ ▷ class-probabilities
    $\ell \leftarrow -\log p[c]$ ▷ cross-entropy loss for target
    $g \leftarrow \nabla_x \ell$ ▷ gradient w.r.t. input
    $x \leftarrow x - \eta g$ ▷ gradient step
    $x \leftarrow \mathrm{clip}(x, -1, 1)$ ▷ keep in valid audio range
end for
return x

- **Initialize:** $x_0 \sim \mathcal{N}(0, I)$
- **Refine:** update x by gradient descent to minimize $\ell_{\mathrm{CE}}(f(x), c)$
- **Clamp:** project x back into the valid audio range after each step
- **Terminate:** stop when $p(c \mid x)$ exceeds a confidence threshold or max iterations reached

Thank you

Thanks for your attention!

References I

[1] AT&T Laboratories Cambridge. *The ORL Database of Faces.* Online: http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html. accessed 2025-05-06.

[2] Matthew Fredrikson, Somesh Jha, and Thomas Ristenpart. "Model inversion attacks that exploit confidence information and basic countermeasures". In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS).* 2015, pp. 1322–1333. DOI: 10.1145/2810103.2813677.

[3] John S. Garofolo. *TIMIT Acoustic-Phonetic Continuous Speech Corpus.* Tech. rep. Linguistic Data Consortium, 1993. URL: https://catalog.ldc.upenn.edu/LDC93S1.

[4] Kevin Pizzi, Florian Boenisch, Utku Sahin, and Klaus Böttinger. "Introducing Model Inversion Attacks on Automatic Speaker Recognition". In: *arXiv preprint arXiv:2301.03206* (2023). URL: https://arxiv.org/abs/2301.03206.

References II

[5] Mirco Ravanelli and Yoshua Bengio. "Speaker recognition from raw waveform with SincNet". In: *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 1021–1028. DOI: 10.1109/SLT.2018.8639541.

[6] Mary-Ann Russon. "Mission: Impossible– Dead Reckoning's technology unpacked— From AI to facial recognition". In: *The Evening Standard* (July 13, 2023). accessed 2025-05-06. URL: https://www.standard.co.uk/news/tech/mission-impossible-dead-reckoning-technology-used-real-life-b1093576.html.

[7] Zhipeng Zhang. *MIA: Model Inversion Attack Toolkit*. GitHub repository. Online: https://github.com/zhangzp9970/MIA (accessed 2025-05-06). 2020.