Opting Out of Facial Recognition

Gavin Taylor
US Naval Academy
Which Stores Are Scanning Your Face? No One Knows.

BankID på mobil blir historie: – Vil gjøre hverdagen enklere

FBI, Pentagon helped research facial recognition for street cameras, drones

Madison Square Garden Uses Facial Recognition to Ban Its Owner’s Enemies
Facial Recognition Embeddings
Dataset of “Gallery Images”

“Probe Image”

Facial Recognition System

Most similar faces
Dataset of “Gallery Images”

“Probe Image”

Facial Recognition System

Most similar faces
Dataset of “Gallery Images”

“Probe Image”

Facial Recognition System

Most similar faces
Dataset of “Gallery Images”

Facial Recognition System

“Probe Image”

Most similar faces
Dataset of “Gallery Images”

Facial Recognition System

“Probe Image”

Most similar faces
GOAL: Manipulate Gallery Image to still serve its purpose, but make it unsuitable for comparison in a black-box facial recognition system.
“Adversarial” examples for ML

Panda, 57.7% confidence  

Gibbon, 99.3% confidence

[Goodfellow et al., “Explaining and Harnessing Adversarial Examples” ICLR 2015]
“Adversarial” examples for ML

Given a neural network, its parameters $\theta$, an image $x$, and a loss function $\mathcal{L}(x, \theta)$, maximize the loss function by altering the image a limited amount ($\|\Delta x\|_\infty < \epsilon$).
Why does this work?
Given a neural network, its parameters $\theta$, an image $x$, and a loss function $L(x, \theta)$, maximize the loss function by altering the image a limited amount ($\|\Delta x\|_\infty < \epsilon$).
Black-box adversarial examples: “Ensemble” approach

Construct several neural networks, and construct adversarial permutations that affect the loss function on all of them - empirically transferable

$$\sum_i L(x, \theta_i)$$
GOAL: Manipulate Gallery Image to still serve its purpose, but make it unsuitable for comparison in a black-box facial recognition system.
LowKey Optimization Function

$$\max_{x'} \frac{1}{2n} \sum_i \frac{\|f_i(x) - f_i(x')\| + \|f_i(x) - f_i(G(x'))\|}{\|f_i(x)\|} - \alpha \text{LPIPS}(x, x')$$

- $x$: Cropped and aligned facial image
- $f_i(x)$: Embedding by model $i$
- LPIPS: Measure of perceptual difference between two image
- $G(x)$: Gaussian-smoothed facial image
Clean Images

Images protected with LowKey
Effectiveness Against Industrial Black Boxes

- 100,000 images from 530 identities, plus 1 million distractor images
- 100 identities randomly chosen, and all images from those identities manipulated
- If any image from that identity appears in the set of possible matches, the facial recognition system has succeeded
Effectiveness Against Industrial Black Boxes

- 100,000 images from 530 identities, plus 1 million distractor images
- 100 identities randomly chosen, and all images from those identities manipulated
- If any image from that identity appears in the set of possible matches, the facial recognition system has succeeded

<table>
<thead>
<tr>
<th></th>
<th>Amazon Rank-1</th>
<th>Amazon Rank-50</th>
<th>Microsoft Rank-1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>93.7%</td>
<td>95.4%</td>
<td>87.7%</td>
</tr>
<tr>
<td>LowKey</td>
<td>0.6%</td>
<td>2.4%</td>
<td>0.1%</td>
</tr>
</tbody>
</table>
Try it yourself: https://lowkey.umiacs.umd.edu/