Private Neural Portraits

Ender Minyard
ML Collective
cum2102@columbia.edu

Abstract

Facial recognition systems learn from datasets whose data subjects have not consented to being subjects.To preserve the privacy of these individuals, researchers have designed systems like LowKey[1] and Fawkes[2] that add adversarial perturbations to images containing faces.These adversarial perturbations make it difficult for facial recognition models to correctly label faces in an image.This work proposes a strategy in which the artistic stylizing of an image containing a face makes it difficult for a facial recognition model to correctly label the face in the image.

1 Motivation

Facial recognition datasets can contain images of individuals who have not consented to being a data subject. Corporations regularly scrape the Internet to increase the size of their facial recognition datasets[3].

The ImageNet dataset originally contained faces that were not obfuscated. ImageNet curators later obfuscated the faces of data subjects out of concern for the privacy of the data subjects[4].

Systems like Fawkes[2] and LowKey[1] let potential data subjects concerned about their privacy obfuscate personal images. However, defenses applied to facial recognition systems like adversarial training[3] and image super-resolution[5] make it difficult for LowKey[1] and Fawkes[2] to protect privacy-concerned users from facial recognition models in the long term.

We propose the use of a Paint Transformer[6] for obfuscating faces in natural images.

We demonstrate the performance of our strategy in a case study. In our case study, Fawkes applies adversarial perturbations to an image. A Paint Transformer stylizes this image. ESRGAN [7] then upscales the image.

MTCNN[8] is a model that detects faces. Face detection is required for face recognition. Facial recognition systems cannot label faces that have not been detected first.

For our experiment, we test the ability of MTCNN[8] to detect faces in the upscaled image.

2 Methodology

The code and data needed to reproduce these results is contained in a Jupyter notebook on GitHub at this url: https://github.com/enderminyard/private-neural-portraits

The data used for our case study is an image of a celebrity from Wikipedia [9]. In our case study, Fawkes obfuscates this image of a celebrity[9]. A Paint Transformer[6] stylizes the image.

ESRGAN[7] upscales both the stylized and unstylized images. MTCNN[8] tries to detect faces in both the stylized and unstylized images.

This experiment was run using an NVIDIA Tesla T4 GPU.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

3 Results

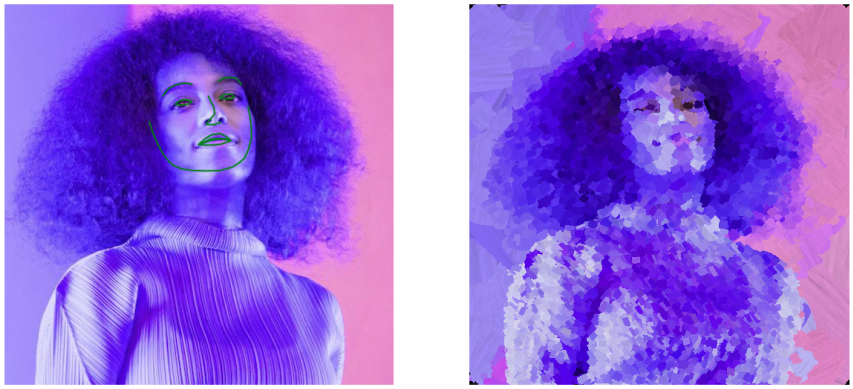

Table 1: MTCNN evalution of perturbed and upscaled images

	Fawkes	Fawkes + Paint Transformer
Confidence	0.9934871196746826	*none*
Bounding box	$[1609, 859, 1039, 1164]$	N/A

In our case study, we find that MTCNN[8] does not detect a face in the upscaled image stylized with the Paint Transformer.

MTCNN[8] is able to detect a face in the upscaled image that is not stylized with a confidence score of 0.9934871196746826.

Fawkes applied perturbations to both images before the image super-resolution process.

4 Discussion

Our results suggest that the Paint Transformer[6] is effective for defeating super-resolution defenses[5] against adversarial perturbations applied to images of faces.

4.1 Further Work

The ability of MTCNN to detect faces in stylized images could change depending on the image being evaluated. In order to conduct a more comprehensive experiment, we need to test the performance of facial recognition systems using a dataset whose subjects have actively consented to being data subjects or consider how to evaluate the performance of a facial detection system without a dataset. Further work in this area may require research in the direction of participatory AI [10].

4.2 Broader Impact

Privacy-concerned individuals aware of becoming potential data subjects without their expressed consent may consider using the strategy described within this work to decrease the probability of being labeled correctly by facial detection systems.

References

[1] Valeriia Cherepanova, Micah Goldblum, Harrison Foley, Shiyuan Duan, John Dickerson, Gavin Taylor, and Tom Goldstein. Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition, 2021. URL https://arxiv.org/abs/2101.07922.

[2] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y. Zhao. Fawkes: Protecting privacy against unauthorized deep learning models. 2020. doi: 10.48550/ARXIV.2002.08327. URL https://arxiv.org/abs/2002.08327.

[3] Evani Radiya-Dixit, Sanghyun Hong, Nicholas Carlini, and Florian Tramèr. Data poisoning won't save you from facial recognition, 2021. URL https://arxiv.org/abs/2106.14851.

[4] Kaiyu Yang, Jacqueline Yau, Li Fei-Fei, Jia Deng, and Olga Russakovsky. A study of face obfuscation in imagenet, 2021. URL https://arxiv.org/abs/2103.06191.

[5] Aamir Mustafa, Salman H. Khan, Munawar Hayat, Jianbing Shen, and Ling Shao. Image super-resolution as a defense against adversarial attacks. *IEEE Transactions on Image Processing*, 29: 1711–1724, 2020. doi: 10.1109/tip.2019.2940533. URL https://doi.org/10.1109%2Ftip.2019.2940533.

[6] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Ruifeng Deng, Xin Li, Errui Ding, and Hao Wang. Paint transformer: Feed forward neural painting with stroke prediction, 2021. URL https://arxiv.org/abs/2108.03798.

[7] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Chen Change Loy, Yu Qiao, and Xiaoou Tang. Esrgan: Enhanced super-resolution generative adversarial networks, 2018. URL https://arxiv.org/abs/1809.00219.

[8] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23:1499–1503, 2016.

[9] Tore Saetre, Jun 2017. URL https://en.wikipedia.org/wiki/Solange_Knowles#/media/File:Solange_(220707).jpg.

[10] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Díaz, Madeleine Elish, Iason Gabriel, and Shakir Mohamed. Power to the people? opportunities and challenges for participatory ai, 09 2022.