Fake Face Detection

Alex Kyllo John Wyman Will Thomas

December 1, 2020

Abstract

In this study we investigate the question of whether it is still feasible to automatically discern AI-generated human face images from genuine photographic ones, by training a convolutional neural network on a labeled dataset of 70,000 real and 70,000 fake face images. We use the fake face recognition problem to further explore the topic of model fairness, by evaluating the model's performance across age, gender and race groups on a demographically labeled face dataset. To achieve this, we propose a method of utilizing an encoder network to translate demographically labeled real face images into an approximation of their latent space representation and then reconstruct them, creating a dataset of matching fake face images with the same demographic labels. This allows us to assess whether our fake face detection model and the generative model that generated its input images, were trained on a demographically biased dataset.

Introduction

Generative Adversarial Networks (GANs) have created the ability to encode photographic images into a latent space representation and automatically generate many images that can appear to be genuine photographs, to the human eye. NVIDIA's StyleGAN [1] model, trained on a dataset of human face images, is capable of generating extremely photo-realistic images of people who do not exist.

An issue with the available

open datasets of human faces is bias in the demographic composition of the pictured individuals.

The FairFace [2] study introduced a new dataset of human face images collected from public datasets with manually verified age, gender and race labels.

While the FairFace dataset provides real human face images that can be used to assess disparities in true negative and false positive rates, a second, similarly labeled dataset of fake face images is needed to compute true positive and false negative rates for specific age, gender and race groups. A research team at Tel Aviv University recently developed a novel encoder network [3] that is capable of approximately reconstructing StyleGAN's latent code representation of a face image and then decoding it back into an image, leading to a fake face output image that closely resembles the real face input image.

Methods

Results

Discussion

References

[1] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *CoRR*, abs/1812.04948, 2018.

[2] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913*, 2019.

[3] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. *arXiv preprint arXiv:2008.00951*, 2020.