Fake Face Detection

Alex Kyllo John Wyman Will Thomas

December 9, 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 machine learning community has recently been struggling with the issue of model fairness–it is important that models perform equitably for users and data subjects of different backgrounds, and also very difficult to enumerate and quantify the sources of bias in training data that can contribute to biased model performance.

The FairFace [2] study introduced a new dataset of human face images collected from public datasets with manually verified, crowd-sourced 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. To address this gap, we investigated methods for "falsifying" a real face image by autoencoding it via the StyleGAN latent space. 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 very closely resembles the real face input image, implying that the original demographic labels would remain valid.

Methods

Data Preprocessing and Augmentation

We tested several methods for preprocessing and augmenting the image data before feeding it into the CNN model.

1. 3-color (RGB) images vs. grayscale
2. Pre-cropping and centering faces using pre-trained face detection models
3. Random horizontal flips

For pre-cropping, we utilized two different pre-trained face detection models, MTCNN and DLib. (TODO: citations)

Model Training

To solve the binary classification task of distinguishing between real and fake human face images, we trained several variations of deep Convolutional Neural Networks (CNN), varying the number of convolution layers as well as several model hyperparameters and image preprocessing steps.

Our baseline model was a CNN with three convolution layers using a 3x3 kernel size

Model Serving

Model Explainability

Model Evaluation

Our primary metric for performance assessment during training and model selection was validation set accuracy, because the balanced classes of the input dataset made accuracy straightforward to interpret. For final model performance on out-of-sample test data, we break down performance with a 2x2 confusion matrix and report F1 score, precision score and recall score in addition to accuracy score.

For fairness metrics, we compare false positive rate and false negative rate ratios for the following binary group definitions taken from the FairFace labels:

1. Gender = “male” compared to Gender = “female”
2. Race = “white” compared to all other races
3. Age = “0-2” compared to all other ages
4. Age = “3-9” compared to all other ages
5. Age = “more than 70” compared to all other ages

We examine the model fairness for children and senior citizens as a recent study [4] found that the popular face recognition model Face++ disproportionately fails to recognize children’s faces in images collected from social media.

Results

Preprocessing

Preprocessing the images using a pre-trained model

Model Performance

TODO: model performance metrics go here

Fairness Assessment

TODO: model fairness metrics go here

Discussion

Future Work

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.

[4] A. Mashhadi, S. G. Winder, E. H. Lia, and S. A. Wood. Quantifying biases in social media analysis of recreation in urban parks. In *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 1–7, March 2020.