On GANs perpetuating biases for face verification

Sasikanth Kotti, Mayank Vatsa, and Richa Singh
{kotti.1,mvatsa,richa}@iitj.ac.in
IIT Jodhpur, India

Abstract. Deep Learning systems need large data for training. Datasets for training face verification systems are difficult to obtain and prone to privacy issues. Synthetic data generated by generative models such as GANs can be a good alternative. However, we show that data generated from GANs are prone to bias and fairness issues. Specifically GANs trained on FFHQ dataset show bias towards generating white faces in the age group of 20-29. We also demonstrate that synthetic faces cause disparate impact, specifically for race attribute, when used for fine tuning face verification systems. This is measured using $DoB_{fv}$ metric, which is defined as standard deviation of GAR@FAR for face verification.

Keywords: Bias, Fairness, GANs, Face Verification, Synthetic Data

1 Introduction

Generative Models such as Generative Adversarial Networks (GANs) [6] are basic building blocks in most of image recognition architectures. The task of face verification [3,8] consists of verifying if the given pair of faces belongs to the same identity. Deep Learning based algorithms for face recognition [13,1] and verification [3,8] utilize face datasets for training. However, obtaining more data is not always easy and even sometimes not possible. GANs can be used to obtain synthetic data where data is scarce and in scenarios where privacy is important. However, existing models (GANs) trained with FFHQ dataset [11] are prone to bias and fairness issues. In this work, we analyze bias and fairness of GANs and their impact on face verification systems.

2 Methodology

Bias and fairness of existing generative models are understood by obtaining attributes and performing domain adaptation on pretrained models. This is specifically understood in the context of face verification task. We describe below the datasets and protocols followed for evaluating bias and fairness.
2.1 Datasets

The Balanced Faces in the Wild (BFW) \cite{12} dataset is balanced across eight subgroups. This consists of 800 face images of 100 subjects, each with 25 face samples. The BFW dataset is grouped into ethnicities (i.e., Asian (A), Black (B), Indian (I), White (W)) and genders (i.e., Females (F) and Males (M)). The metadata for this dataset consists of a list of pairs for face verification. Hence, this dataset can be used to investigate bias in automatic facial recognition (FR) system for verification.

CMU Multi-PIE \cite{7} is a constrained dataset that consists of over 44,000 images of 336 subjects. These are frontal faces with variation in illumination and expressions.

FFHQ, which stands for Flickr-Faces-HQ \cite{11} is a dataset of 70,000 human faces of high resolution 1024x1024 and covers considerable diversity and variation. Images with permissive license were obtained from Flickr.

2.2 Evaluation Protocol

Evaluation for estimation of bias and fairness is performed in two phases. Initially, the proportion of faces generated for each sub-group of different attributes such as Age, Gender, Race and Race4 were analysed. In the next phase, a pre-trained face verification model is fine-tuned, and the impact of fairness is analyzed using Degree Of Bias (DoB) metric. We define the Degree Of Bias (DoB) for face verification as the standard deviation of GAR@FAR:

\[
DoB_{fv} = \sqrt{\frac{\sum (GAR_{sg} - \mu)^2}{N}}
\]

where
- \(GAR_{sg}\) = GAR @ FAR for each sub-group
- \(\mu\) = Mean GAR@FAR
- \(N\) = Number of sub-groups
- GAR = Genuine Accept Rate
- FAR = False Accept Rate

3 Experiments

We describe below the experimental setup, results and our findings.

3.1 Experimental Setup

The generator of StyleGAN2 with adaptive discriminator augmentation (ADA) \cite{10} trained on the FFHQ dataset is used to generate synthetic face images. The attributes such as race, race4, gender and age of these synthetic faces were obtained using a pretrained Fairface \cite{9} attribute classifier. The proportion of images for each attribute type was plotted, to understand bias in synthetic faces for each
attribute. For the next section, DiscoFaceGAN [5] is considered for generating different faces for different identities, expressions, lighting and poses. VGGFace2 [2] model is considered for domain adaptation with CMU Multi-PIE [7] and synthetic faces generated with DiscoFaceGAN [5]. About 10000 synthetic faces of 2500 identities were generated with DiscoFaceGAN [5]. Out of these, 2000 identities were used for training and, 500 identities were used for validation. The 336 subjects of CMU Multi-PIE [7] were split into 70-30 ratio for training and validation. Fine-tuning was carried out for 10 epochs with a learning rate of 1e-4, batch size of 128, weight decay of 1e-4 and momentum of 0.9. The last two convolutional layers of VGGFace2 [6] were fine-tuned with ArcFace [4] loss of margin 35 and scale 64. The checkpoint with the lowest validation loss is considered for inference with BFW dataset [12]. Inference is carried out by using Cosine distance between the pairs. Comparison of GAR@FAR for different attributes such as race, gender and others is carried out for bias estimation.

3.2 Results

From figure [1] it is evident that GANs trained with the FFHQ dataset are biased towards generating faces in the age group "20-29" and mostly "White" faces. However, no bias is observed for gender attribute. Figure [2] demonstrates
4 Conclusion

GANs are popular networks that are very successful in generating faces of good perpetual quality. These are trained with existing datasets. However, the biases present in the dataset are also being manifested in these networks. We analyzed the biases of these networks for important attributes such as Age, race and gender for faces. We also demonstrated how this could impact the sub-group performance of downstream models such as face verification systems. Hence, it is important to debias GANs before using them in any application. In future, we aim to investigate methods and techniques for debiasing GANs with respect to different critical attributes.

Fig. 2. $DoB_{fv}$ i.e Std(GAR @ FAR) for Ethnicity, Gender and Attributes with CMU Multi-Pie and Synthetic faces (smaller is better for bias)

that Face Verification models, when fine-tuned with Synthetic data generated by GANs exhibit bias towards race. This is measured by comparing the $DoB_{fv}$ for models fine-tuned with CMU MultiPie and Synthetic faces. $DoB_{fv}$ is greater for models trained with Synthetic faces. This is predominant at low FAR rates. For, high FAR rates we don’t observe bias. We hypothesize that although biases are present these are masked by high false acceptances. Bias for gender attribute is not observed in Synthetic faces, hence no gender bias is manifested by fine-tuned face verification systems.
References

1. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Transactions on pattern analysis and machine intelligence 19(7), 711–720 (1997)
2. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age (2018)
3. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05). vol. 1, pp. 539–546. IEEE (2005)
4. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition (2019)
5. Deng, Y., Yang, J., Chen, D., Wen, F., Tong, X.: Disentangled and controllable face image generation via 3d imitative-contrastive learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
6. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems 27 (2014)
7. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. Image and vision computing 28(5), 807–813 (2010)
8. Huang, G.B., Lee, H., Learned-Miller, E.: Learning hierarchical representations for face verification with convolutional deep belief networks. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 2518–2525. IEEE (2012)
9. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. In: Proc. NeurIPS (2020)
10. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
11. Robinson, J.P., Livitz, G., Henon, Y., Qin, C., Fu, Y., Timoner, S.: Face recognition: too bias, or not too bias? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–1 (2020)
12. Turk, M.A., Pentland, A.P.: Face recognition using eigenfaces. In: Proceedings. 1991 IEEE computer society conference on computer vision and pattern recognition. pp. 586–587. IEEE Computer Society (1991)
A Detailed Description and Visualization of Datasets, Architectures and Results

A.1 Datasets
Figure 3 shows CMU Multi-PIE [7] which is a constrained dataset and FFHQ, which stands for Flickr-Faces-HQ [11].

Figure 4 shows Balanced Faces in the Wild (BFW) [12] which are used for evaluating bias for face verification task and Synthetic Faces generated with DiscoFaceGAN [5].

A.2 Architectures
Attributes such as race, gender and age for synthetic faces generated by GAN are obtained using a pretrained Fairface [9] attribute classifier. The proportion of
images for each attribute are analyzed for imbalance and bias. This architecture is shown in Figure 5.

![Fig. 5. GAN Bias Estimation Architecture](image)

As shown in Figure 6 the impact of bias and fairness on face verification systems is analyzed by fine-tuning with CMU Multi-PIE [7] and Synthetic Faces generated with DiscoFaceGAN [5] and comparing $DoB_{fv}$ i.e Std(GAR @ FAR) for different attributes.

![Fig. 6. Bias Estimation in Face Verification System](image)

### A.3 Results

Table 1 shows GAR@FAR when face recognition model was fine-tuned with CMU Multi-PIE and Synthetic faces. The overall performance is similar for both the datasets. Table 2 and Table 3 shows GAR@FAR and their standard deviations ($DoB_{fv}$) for each sub-group of gender and ethnicity attributes.
| GAR(%) | CMU Multi-PIE | Synthetic Faces |
|--------|---------------|-----------------|
| 0.01   | 21.59         | 22.77           |
| 0.1    | 38.45         | 39.51           |
| 1      | 62.61         | 63.07           |
| 10     | 88.02         | 88.05           |

**Table 1.** GAR@FAR

| GAR(%) | CMU Multi-PIE | Synthetic Faces |
|--------|---------------|-----------------|
| 0.01   | 22.77         | 19.98           |
| 0.1    | 41.47         | 36.6            |
| 1      | 66.23         | 60.55           |
| 10     | 88.85         | 87.54           |

**Table 2.** GAR@FAR for gender attribute

| GAR(%) | CMU Multi-PIE | Synthetic Faces |
|--------|---------------|-----------------|
| 0.01   | 16.2          | 22.24           |
| 0.1    | 30.05         | 36.57           |
| 1      | 52.46         | 66.52           |
| 10     | 82.38         | 87.85           |

**Table 3.** GAR@FAR for ethnicity attribute