

On Biased Behavior of GANs 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 biased behavior 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.

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

1 Introduction

Generative Models such as Generative Adversarial Networks (GANs) [9,29,2] are basic building blocks in most of image recognition architectures. The task of face verification [6,12,25,21] consists of verifying if the given pair of faces belongs to the same identity. Deep Learning based algorithms for face recognition [26,3,1,19] and verification [6,12,25,21] 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 [17] are prone to bias and fairness issues. In this work, we analyze bias and fairness of GANs [18] and their impact on face verification systems. Our main contributions in this research are as follows :

- **Result-1:** We observed that GANs trained on FFHQ dataset exhibit bias for the "age" and "race" protected attributes.
- **Result-2:** We demonstrate that Face Verification systems that are trained or fine-tuned with GAN data exacerbate bias for the "race" protected attribute.

2 Datasets and Protocol

In this section we briefly describe the datasets and evaluation protocol. In the experiments, we use three datasets:

- The Balanced Faces in the Wild (BFW) [24] 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)).
- CMU Multi-PIE [10] is a constrained dataset consisting of face images of 337 subjects with variation in pose, illumination and expressions. Of these over 44K images of 336 subjects images are selected corresponding to frontal face images having illumination and expression variations.
- FFHQ or Flickr-Faces-HQ [17] is a dataset of 70,000 human faces of high resolution 1024x1024 and covers considerable diversity and variation.

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 pretrained face verification model is fine-tuned, and the impact of fairness is analyzed using Degree Of Bias(DoB) metric. We define DoB for face verification as the standard deviation of GAR@FAR

$$DoB_{fv} = \sqrt{\frac{\sum (GAR_{sg} - \mu)^2}{N}} \quad (1)$$

where GAR_{sg} stands for GAR @ FAR for each sub-group, μ represents mean of GAR@FAR and N represents number of sub-groups. The GAR and FAR stands for Genuine Accept Rate and False Accept Rate respectively.

3 Experiments

Experiment-1: As part of this experiment the generator of StyleGAN2 with adaptive discriminator augmentation (ADA) [16] trained on the FFHQ dataset is used to generate synthetic face images. Attributes such as race, race4, gender and age of these generated synthetic faces were obtained using a pretrained Fairface [15] attribute classifier. The proportion of images for each attribute type were plotted, to understand bias and imbalance.

Experiment-2: In this experiment DiscoFaceGAN [8] is considered for generating different faces for different identities, expressions, lighting and poses. VGGFace2 [4] model is considered for domain adaptation with CMU Multi-PIE [10] and synthetic faces generated with DiscoFaceGAN [8].

About 10000 synthetic faces of 2500 identities were generated with DiscoFaceGAN [8]. Out of these, 2000 identities were used for training and, 500 identities were used for validation. Similarly the 336 subjects of CMU Multi-PIE [10] were split into 70-30 ratio for training and validation. Fine-tuning with both datasets 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 [7] loss of margin 35 and scale 64.

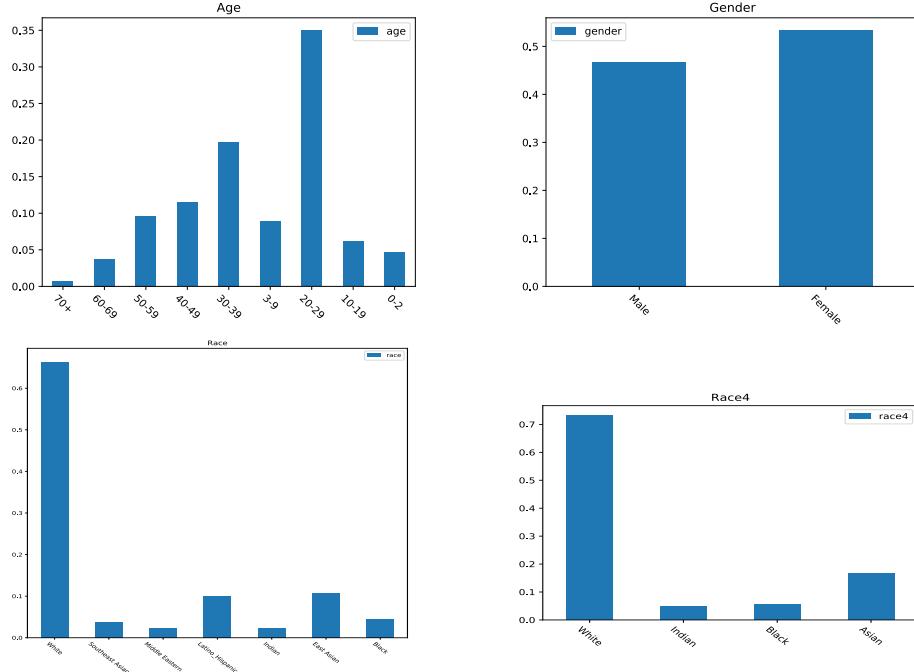


Fig. 1. Proportion of faces in each sub-group for Age,Gender,Race and Race4 attribute for GAN generated synthetic faces(x-axis sub-groups and y-axis proportion)

The checkpoint with the lowest validation loss for each dataset is considered for inference with BFW dataset [24].

Inference is carried out by using Cosine distance between the pairs. Comparison of DoB_{fv} i.e Std(GAR @ FAR) for different attributes such as race, gender and others is carried out for both models.

4 Results and Analysis

From figure 1, it is evident that GANs trained with the FFHQ dataset are biased towards generating more faces in the age group "20-29" and mostly "White" faces. However, no such imbalance is observed for gender attribute.

Figure 2 shows the performance of Face Verification models for different face attributes such as Ethnicity, Attributes and Gender. Bias and fairness 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. This behaviour is not observed at high FAR rates. Our observations from the analysis of the results are as follows : (Observation-1 is drawn from experiment-1 and observations-2,3,4 were drawn from experiment-2)

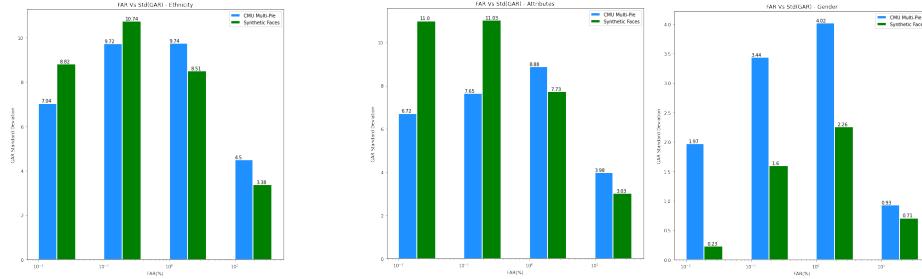


Fig. 2. DoB_{fv} i.e Std(GAR) @ FAR for Ethnicity, Gender and Attributes with CMU Multi-Pie and Synthetic faces (smaller is better for bias)

- **Observation-1:** GANs are biased towards age group "20-29" and "White" faces.
- **Observation-2:** Face Verification models trained or fine-tuned with Synthetic faces exhibit bias for "race" attribute. This is confirmed by high DoB_{fv} for Synthetic faces when compared to CMU MultiPie.
- **Observation-3:** Face Verification models trained or fine-tuned with Synthetic faces doesn't exhibit any bias for "gender" attribute.
- **Observation-4:** At, high FAR rates we don't observe bias (low DoB_{fv}). We hypothesize that although biases are present these are masked by high false acceptances.

5 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.

References

1. An, X., Deng, J., Guo, J., Feng, Z., Zhu, X., Yang, J., Liu, T.: Killing two birds with one stone: Efficient and robust training of face recognition cnns by partial fc. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4042–4051 (2022)
2. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International conference on machine learning. pp. 214–223. PMLR (2017)
3. 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)
4. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age (2018)
5. Chang, J.R., Chen, Y.S., Chiu, W.C.: Learning facial representations from the cycle-consistency of face. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9680–9689 (2021)
6. 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)
7. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)
8. 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)
9. 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)
10. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. *Image and vision computing* **28**(5), 807–813 (2010)
11. Hou, X., Li, Y., Wang, S.: Disentangled representation for age-invariant face recognition: A mutual information minimization perspective. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3692–3701 (2021)
12. 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)
13. Huang, Y., Wu, J., Xu, X., Ding, S.: Evaluation-oriented knowledge distillation for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18740–18749 (2022)
14. Karakas, C., Dirik, A., Yalcinkaya, E., Yanardag, P.: Fairstyle: Debiasing stylegan2 with style channel manipulations (2022). <https://doi.org/10.48550/ARXIV.2202.06240>, <https://arxiv.org/abs/2202.06240>
15. Karkkainen, K., Joo, J.: Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1548–1558 (2021)
16. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. In: Proc. NeurIPS (2020)

17. 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)
18. Lang, O., Gandelsman, Y., Yarom, M., Wald, Y., Elidan, G., Hassidim, A., Freeman, W.T., Isola, P., Globerson, A., Irani, M., et al.: Explaining in style: Training a gan to explain a classifier in stylespace. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 693–702 (2021)
19. Liu, C., Yu, X., Tsai, Y.H., Faraki, M., Moslemi, R., Chandraker, M., Fu, Y.: Learning to learn across diverse data biases in deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4072–4082 (2022)
20. Liu, J., Qiu, D., Yan, P., Wei, X.: Learn to cluster faces via pairwise classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3845–3853 (2021)
21. Lu, C., Tang, X.: Surpassing human-level face verification performance on lfw with gaussianface. In: Twenty-ninth AAAI conference on artificial intelligence (2015)
22. McDuff, D., Ma, S., Song, Y., Kapoor, A.: Characterizing bias in classifiers using generative models. arXiv preprint arXiv:1906.11891 (2019)
23. Park, S., Lee, J., Lee, P., Hwang, S., Kim, D., Byun, H.: Fair contrastive learning for facial attribute classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10389–10398 (2022)
24. 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)
25. Sun, Y., Wang, X., Tang, X.: Hybrid deep learning for face verification. In: Proceedings of the IEEE international conference on computer vision. pp. 1489–1496 (2013)
26. 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)
27. Wang, Z., Dong, X., Xue, H., Zhang, Z., Chiu, W., Wei, T., Ren, K.: Fairness-aware adversarial perturbation towards bias mitigation for deployed deep models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10379–10388 (2022)
28. Zhang, Y., Deng, W., Zhong, Y., Hu, J., Li, X., Zhao, D., Wen, D.: Adaptive label noise cleaning with meta-supervision for deep face recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15065–15075 (2021)
29. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)
30. Zhu, W., Wang, C.Y., Tseng, K.L., Lai, S.H., Wang, B.: Local-adaptive face recognition via graph-based meta-clustering and regularized adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20301–20310 (2022)

A Detailed Description and Visualization of Datasets, Architectures and Results

A.1 Datasets

Figure 3 shows CMU Multi-PIE [10] which is constrained dataset and FFHQ, which stands for Flickr-Faces-HQ [17]



Fig. 3. CMU Multi-PIE and FFHQ Datasets

Figure 4 shows Balanced Faces in the Wild (BFW) [24] which are used for evaluating bias for face verification task and Synthetic Faces generated with DiscoFaceGAN [8]



Fig. 4. BFW and Synthetic(DiscoFaceGAN) Faces

A.2 Architectures

Attributes such as race, gender and age for synthetic faces generated by GAN are obtained using a pretrained Fairface [15] attribute classifier. The proportion of

images for each attribute are analyzed for imbalance and bias. This architecture is shown in Figure 5

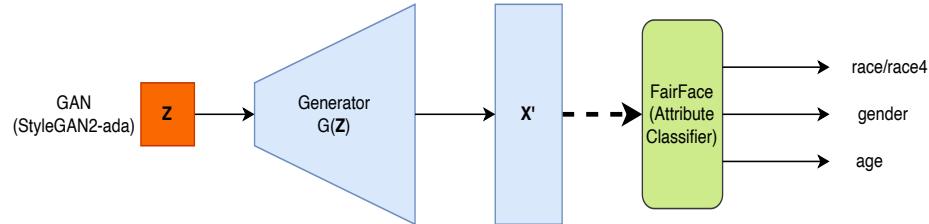


Fig. 5. GAN Bias Estimation Architecture

As shown in Figure 6 the impact of bias and fairness on face verification systems is analyzed by fine-tuning with CMU Multi-PIE [10] and Synthetic Faces generated with DiscoFaceGAN [8] and comparing DoB_{fv} i.e Std(GAR @ FAR) for different attributes

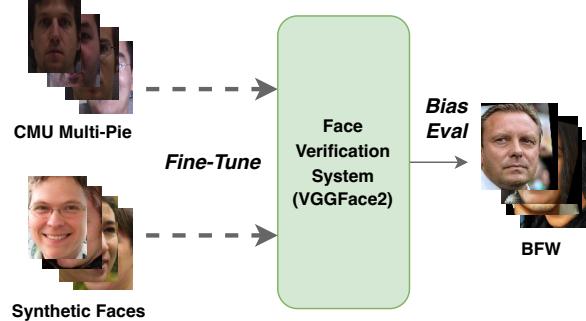


Fig. 6. Bias Estimation in Face Verification System

A.3 Results

Table 1 shows GAR@FAR when face recognition model was fine-tuned with CMU Muti-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(%)		
FAR(%)	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(%)		
FAR(%)	CMU Multi-PIE	Synthetic Faces
	Male(M)	Female(F)
0.01	22.77	19.98
0.1	41.47	36.6
1	66.23	60.55
10	88.85	87.54
	1.97	22.71
	23.04	0.23
	38.36	1.60
	61.43	2.26
	87.53	0.71

Table 2. GAR@FAR for gender attribute

GAR(%)										
FAR(%)	CMU Multi-PIE					Synthetic Faces				
	Asian(A)	Black(B)	Indian(I)	White(W)	Std	Asian(A)	Black(B)	Indian(I)	White(W)	Std
0.01	16.2	22.24	24.37	31.19	7.04	18.0	18.53	27.5	36.69	8.82
0.1	30.05	38.27	43.57	53.23	9.72	31.35	36.75	44.48	56.11	10.74
1	52.46	62.25	65.52	76.08	9.74	56.55	60.08	64.95	76.10	8.51
10	82.38	88.09	87.85	93.4	4.50	84.5	86.86	88.06	92.56	3.38

Table 3. GAR@FAR for ethnicity attribute