Date: March 31, 2024

To: Amy Pavlov, CTO

From: Heami Oh, Research Engineer

Subject: Recommendation on Method for Developing Facial Recognition Without Bias

FOREWORD

Our company is looking for an effective method of reducing and eliminating bias in the facial recognition system we are currently developing. Accordingly, I was asked to investigate bias in facial recognition systems and make a recommendation for the optimal approach to reducing bias. The purpose of this report is to provide a recommendation for a methodology that will minimize bias and aid in creating an ethical facial recognition system.

SUMMARY

I recommend that we use diverse demographics in our training datasets, implement a debiasing approach to handle algorithmic bias, and use deep learning algorithms in our facial recognition system. Studies show that when image recognition technologies are trained on data that is lacking in diversity, the resulting technology is significantly less equipped for identifying minority groups in the data (Grother et al., 2019). Subsequently, this leads to negative consequences when that biased technology is used in real life contexts. Another study outlines a methodology for sourcing and cultivating unbiased training data in order to address this problem (Zhang & Deng, 2020). Yet another option is to build debiasing algorithms that minimize the impact of bias in training datasets. These strategies can be supplemented by using a deep learning approach to avoid common issues of bias that surround more traditional approaches to facial recognition technologies. I suggest we follow this combination of methodologies in order to have success in eliminating bias from our datasets and final system. By following these techniques and incorporating an equal diversity of demographics in our training datasets, we can significantly reduce and potentially eliminate these issues and be better equipped for creating a unbiased technology.

INTRODUCTION

Historically, facial recognition technology (FRT) has been plagued by issues with bias. For example, an investigative report into FRT bias by the National Institute of Standards and Technology (NIST) found that African American, Asian, and American Indian faces had the highest rates of being misidentified across the board for different FRTs; additionally, incorrect matches between images occurred at rates 2–5 times higher for women than for men (Grother et al., 2019). As we can see from these data, it is members of underrepresented groups who are most impacted by inaccuracies in these technologies. As a result, when FRT systems are put to use in civil or criminal contexts, these biases can result in detrimental consequences for these implicated groups.

The goal of my research was to improve these issues in our company's own FRT model by investigating the importance of, as well as different methods for, mitigating bias in FRT. First, I sought to highlight the pressing need for these methods in our own FRT model by investigating the causes and consequences of bias as existing in current usages of FRT. Then, I looked into methods of improving training datasets, specifically focusing on data labelling and sourcing external audits for training datasets. Following this, I investigated a method of targeting algorithmic bias by incorporating debiasing techniques in the algorithm itself. Finally, I researched the FRT methodology followed by the bioinformatics company Idemia, which has been largely successful in building unbiased FRT.

BODY

Section 1: Rationale

The presence of bias in FRT algorithms likely results from training data having an unequal distribution of races and genders, with women and people of color having smaller proportions of images in the datasets. Interestingly, the NIST study notes that algorithms that were developed in China are significantly more accurate with East Asian faces (Grother et al., 2019). This lends further credence to the idea that the performance of FRT models is highly dependent on the makeup of the training datasets used: higher representation in training data translates to higher accuracy in identification for the resulting model.

As one may imagine, the issues in bias with facial recognition technologies can have severe consequences on individuals, particularly those from underrepresented groups. In the article by Gentzel (2020), the author points out that "FRT programs used by law enforcement in identifying crime suspects are substantially more error-prone on facial images depicting darker skin tones and females as compared to facial images depicting Caucasian males," which often results in unfair treatment of people from those groups. As an example, Gentzel (2020) brings up the wrongful arrest of Robert Williams in June 2020: Williams, after being mistakenly identified by a racially biased FRT being used by law enforcement, was wrongly accused by police and

detained for 30 hours before being released on bail, and the case was not dropped until significantly later. It is evident that such cases would have incredibly negative consequences on individuals: not only does it lead to unjust restrictions being put on an innocent person, but it also causes unnecessary emotional and psychological stress that would otherwise have not occurred if biased FRT had not been used in the system. With that in mind, it comes as no surprise that cities such as San Francisco and Boston have banned FRT from being used by city officials or law enforcement, citing the rampant bias present in FRT as a crucial reason (Lunter, 2020).

Keeping these consequences of bias of mind, it is of utmost importance for us to prioritize the reduction or—ideally—elimination of bias in the FRT system we are currently developing. If we do not rigorously cut down bias in our model, it may have detrimental effects on the people it is used on. This will also cut stakeholder support, and, depending on the severity of the issues, it may stall the project entirely. Bias is a critical issue with technology, particularly with FRT, such that it absolutely cannot be overlooked when designing and creating systems.

Section 2: Training Datasets

The quality—or lack thereof—of training datasets is directly reflected in the technologies that train on them. In other words, the performance of our FRT program will be dependent on the quality of the training data we use for it. Accordingly, Lunter (2020) makes the argument that poorly-labelled datasets result in higher rates of misidentification for people of implicated groups, and these misidentifications are learned as correct associations by the FRT system being trained on such data. Subsequently, these accuracy issues can snowball as the algorithm repeatedly trains itself across problematic datasets, resulting in a poorly-performing, biased FRT.

To avert these issues, Lunter (2020) points out that it is important to source robust datasets that have been audited for quality and diversity. One such dataset is BUPT-CBFace, a medium-scale FRT dataset containing 500,000 images that has been shown to effectively reduce bias (Zhang & Deng, 2020). Most FRT training datasets have an unequal distribution of demographics that results in biased algorithms, but BUPT-CBFace overcomes this issue of imbalanced data by incorporating a rich variety of images, with diverse representation, angles, lighting, and expressions (Zhang & Deng, 2020). This results in a remarkably well-balanced dataset, with far more equal distribution of different races and genders.

FRT systems trained on these balanced datasets are able to better identify faces regardless of factors such as race or gender: BUPT-CBFace was shown to outperform much larger datasets, obtaining higher accuracy and lower bias across different races (Zhang & Deng, 2020). This shows how balanced training sets may be utilized to further our goal of building an unbiased FRT model. Another balanced training set is FairFace, which contains 108,501 face images with an equal distribution of races (Kärkkäinen & Joo, 2021). By training our model on unbiased datasets such as BUPT-CBFace and FairFace, we can make large strides toward eliminating bias in our FRT system. It must be noted, however, that these datasets are limited in number, and the

overwhelming majority of publicly available datasets used for FRT are biased in race and gender. While we should use these balanced datasets where possible, we will have to incorporate additional, alternate methods that do not rely on unbiased datasets in order to establish a more universal foundation for reducing bias.

Section 3: Algorithms

Since it may not be feasible to source unbiased training datasets on a large scale, an alternative would be to build robust algorithms that are resilient to bias in training datasets. An example is given in the paper by Amini et al. (2019), which describes methodology for an FRT algorithm that debiases imbalanced data as it trains. This strategy can be briefly summarized as follows: the algorithm first identifies overrepresented categories in data, then adjusts accordingly by adaptively resampling the dataset (Amini et al., 2019). By minimizing these overrepresented features, this approach results in a balanced distribution across different categories in the final sampled training data. In experimental results, bias was reduced by a factor of 3 (Amini et al., 2019). The results of this are shown in Figure 1 below.

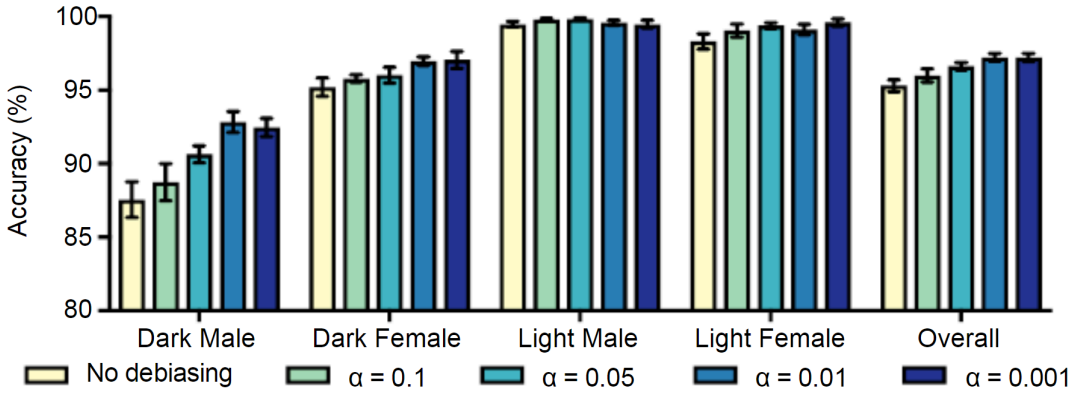

Figure 1. Results of debiasing algorithm, categorized by skin tone and gender. Overall, accuracy increased. (Source: Amini et al., 2019).

As seen in Figure 1 above, the debiasing algorithm increased the overall accuracy in identifying faces. This improvement was most substantial for dark-skinned males; however, the final accuracy is still noticeably below that of light-skinned males. Amini et al. (2019) addresses this shortcoming by explaining that despite the debiasing of training data, the initial data's lower numbers of images of darker-skinned males limits the model in the debiasing process, resulting in these imperfect results. Despite this, the improvement in accuracy for dark-skinned males is significant when compared to other categories, and we can see from the figure that the final

accuracies after debiasing are more balanced across categories of skin tone and gender than when there is no debiasing in the algorithm.

Following a similar methodology may be a strong option for debiasing training datasets and thus removing bias from our resulting FRT program. My recommendation is to ideally use unbiased training datasets (as described in Section 2) in tandem with a bias-resistant algorithm; this will address the problem of limited images for minority groups, and strengthen the debiasing capabilities of our algorithm. Additionally, if we are struggling to source large enough unbiased training datasets, then the debiasing capabilities of our bias-resistant algorithm should substantially help in reducing impacts of bias from the imbalanced training data. In this case, we will likely still end up with a biased model, however, and so further investigation into other approaches is required.

Section 4: Deep Learning & Idemia

In this section, I focus on Idemia's successful approach to unbiased FRT. Idemia is a biometrics company that was noted in NIST's testing to have unbiased FRT, with its algorithm ranking highly in accuracy across all tests (Grother et al., 2019). Rather than pursuing the traditional approach to FRT, where the algorithms focus on specific features of the image to identify and match faces, Idemia uses deep learning algorithms for their FRT. Idemia's Chief AI Scientist Stephane Gentric (2018) points out that because the traditional approach mainly focuses on individual facial features when processing images, it "concentrates the discriminative information". This exaggerates the impact of individual features, which results in identifications being made based on limited information and facilitates bias in the resulting technology. To avoid this, Idemia uses the strategy of building a template of the face from a given image, then identifying people based on matching the facial templates (Gentric, 2018). To aid with this process, they use a deep network to learn and train itself to build accurate, effective facial templates (Gentric, 2018). Since deep learning models rely heavily on loss functions to self-evaluate accuracy as the model trains itself, it is crucial to utilize a loss function that results in the highest accuracy for the given context. For their FRT model, Idemia uses the Von Mises-Fisher Loss function, which enables rebalancing of databases that are imbalanced in qualities such as race, gender, or age (Gentric, 2018).

Idemia's approach has resulted in algorithms that demonstrate little to no bias across different demographics (Gentric, 2018). I suggest we also follow the strategy of using deep learning models to template and match faces, employing the same or a similar loss function in order to control imbalances in training data. This strategy should significantly help us reduce and eliminate bias in our algorithms, which will increase the overall accuracy of our FRT system. I recommend that we use this technique in conjunction with the methods described in Sections 2 and 3; in this way, we will be significantly better equipped to address bias on all levels: in the initial training data, in the underlying algorithm, and in the final model.

CONCLUSION

My investigation concluded that the best approach to mitigating bias in FRT would be an approach that synthesizes the methods outlined in Sections 2–4. I recommend for smaller-scale FRT models, we use unbiased datasets such as BUPT-CBFace to train our models on. For larger-scale models, however, this may not be a feasible option, since there are limited options for unbiased datasets. In both cases, I recommend that we employ bias-resistant algorithms such as the one described in the paper by Amini et al. (2019), which was able to balance uneven distributions in training datasets. We can use this strategy in tandem with the methodology followed by Idemia in their FRT systems: by using the same framework of building facial templates from images and then using deep networks to compare templates to make an identification, we can reduce the emphasis on individual features in identification and thereby reduce the bias that results from such an approach (Gentric, 2018). Following Idemia's strategy, we should also use the Von Mises-Fisher Loss function or a loss function that has similar or better results, in order to mitigate the effects of bias in the underlying datasets. In this manner, we can address bias in both the training data and the created algorithm, which will significantly reduce bias in the resulting FRT model.

It is important to note that this report focused on the initial outline of our next steps, so certain avenues were not explored that otherwise might benefit our FRT design. These unexplored directions include alternate strategies to reduce algorithmic bias, research into other successful examples of unbiased FRT, and testing alternate loss functions to be used with a deep network. In the future, we could implement and test different loss functions with our deep network to see which would be most effective in reducing bias. Furthermore, this report looked into only one method of reducing bias in the algorithm itself. By researching and implementing prototypes of other debiasing methods in the future, and subsequently comparing the different efficacies of reducing bias, we could better determine the most optimal algorithm. Potentially, we could integrate strategies from different methodologies to create an combinatory algorithm best equipped to mitigate bias in training data.

Additionally, this report only looked into the example of Idemia as a template for success; however, other companies have also created FRT models successful in handling bias. One such algorithm is NEC-3, which is described in the NIST report as "the most accurate [they] have evaluated"; the NIST report also mentions developers with higher performing FRT models, such as Aware, Toshiba, Tevian and Real Networks (Grother et al., 2019). By looking into other unbiased FRT algorithms such as NEC-3 and those from other successful developers, we could expand on the recommendations made in this report by further modifying our strategy of cultivating an unbiased FRT algorithm. If we decide to follow the recommendations outlined in this report, I suggest more thoroughly looking into optimal loss functions, methods of implementation for reducing algorithmic bias, and alternate FRT model techniques by successful developers. If these steps are followed, I believe that bias in our FRT model would be

significantly reduced, and an unbiased FRT model would be feasible for us to to achieve.

REFERENCES

Amini, A., Soleimany, A., Schwarting, W., Bhatia, S., & Rus, D. (2019). Uncovering and mitigating algorithmic bias through learned latent structure. *2019 AAAI/ACM Conference*. https://doi.org/10.1145/3306618.3314243

Gentric, S. (2018). Face recognition evaluation @ Idemia. *Proc. International Face Performance Conference*. Retrieved from https://nigos.nist.gov/ifpc2018/presentations/44_gentric_Idemia_IFPC.pdf

Gentzel, M. (2021). Biased face recognition technology used by government: a problem for liberal democracy. *Philosophy & Technology, 34*(4), 1639–1663. https://doi.org/10.1007/s13347-021-00478-z

Grother, P., Ngan, M., & Hanaoka, K. (2019). Face recognition vendor test part 3: demographic effects. *National Institute of Standards and Technology, 8280*. https://doi.org/10.6028/NIST.IR.8280

Kärkkäinen, K., & Joo, J. (2021). FairFace: face attribute dataset for balanced race, gender, and age. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1547-1557. https://doi.org/10.1109/WACV48630.2021.00159

Lunter, J. (2020). Beating the bias in facial recognition technology. *Biometric Technology Today*, *2020*(9), 5–7. https://doi.org/10.1016/S0969-4765(20)30122-3

Zhang, Y., & Deng, W. (2020). Class-balanced training for deep face recognition. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 3594-3603. https://doi.org/10.1109/CVPRW50498.2020.00420