Exploring Racial Bias in Classifiers for Face Recognition

Jaeju An
Sungkyunkwan University
Seongnam-Si, Gyeonggi-do, South Korea
anjaeju@g.skku.edu

Bosung Yang
Sungkyunkwan University
Suwon-Si, Gyeonggi-do, South Korea
qhtlda1@g.skku.edu

Jeongho Kim
Sungkyunkwan University
Suwon-Si, Gyeonggi-do, South Korea
rlawjdghek@g.skku.edu

Geonwoo Park
Sungkyunkwan University
Suwon-Si, Gyeonggi-do, South Korea
rjsdn1120@g.skku.edu

Simon S. Woo
Sungkyunkwan University
Suwon-Si, Gyeonggi-do, South Korea
swoo@g.skku.edu

ABSTRACT

Recent advancements in deep learning have allowed, among others, various applications of face recognition systems, where a large amount of face image data are typically required for training. In addition to the size of the dataset, its composition has a notable impact on the face recognition accuracy of deep learning-based algorithms. This may render several high-performing classifiers inaccurate due to the lack of fairness regarding their training data. Such severe performance drop poses serious problems, especially when biased results are introduced in many data-critical and sensitive applications. In this work, we present a preliminary case study to examine the effect of unfairly composed datasets on popular classifiers that can distinguish genders by training them on imbalanced data in terms of race. We empirically demonstrate that the gender classification accuracy may vary from 3% to as high as 23% due to the ratio of samples corresponding to different races, i.e., Western and Asian, in the training and testing datasets.

CCS CONCEPTS

• **Information systems** → **Data mining**; • **Computing methodologies** → **Computer vision**.

KEYWORDS

Fairness, Racial Bias, Gender Detection

ACM Reference Format:
Jaeju An, Bosung Yang, Jeongho Kim, Geonwoo Park, and Simon S. Woo. 2021. Exploring Racial Bias in Classifiers for Face Recognition. In *Companion Proceedings of the Web Conference 2021 (WWW '21 Companion), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/XXXXXX.XXXXXX

1 INTRODUCTION

Significant advancements in deep learning have contributed to solving the most challenging problems in our society for a wide range of applications, including face recognition. However, the performance of most deep learning algorithms, regardless of the domain, heavily depends on the characteristics of the training data [4]. If the training dataset has an unfair composition, then the corresponding test results may be severely biased and the underlying analyses or implications could also be erroneous. For example, face recognition systems are used for many critical applications, such as user authentication or human identification in crime scenes.

In either of them, both false positives and false negatives due to mediocre classification accuracy lead to serious issues. Hence, the training dataset must contain diverse samples with reasonably balanced ratio. In fact, Merler et al. [10] showed that the most commonly used facial dataset[6, 9] have an unfair data composition in terms of race, with over 80% of the samples representing white people, while only a small portion represents black people. Such lack of fairness in data can lead to biased results and incorrect predictions when, for instance, using face analysis program [1].

In this preliminary study, we develop simple gender classifiers using popular neural network structures, i.e., Xception [2] and MobileNet-v2 [12], to examine the effect of racial bias on the classification performance as an illustrative use case. Our main research question is whether a model trained only on data representing Western men and women can classify Asian men and women as effectively as Western men and women. To that end, we manipulate our dataset such that all training samples solely correspond to Western men and women, and all testing samples to Asian men and women.

We also investigate the effect of biased data on fine-tuned deep learning classifiers. Through our experimental results, we clearly demonstrate that models trained only on data representing Western people result in lower accuracy when detecting Asian people, while fine-tuned networks from pre-trained ones result in accuracy increase of up to over 20% when detecting Western people; we find that the resulting skewed performance also holds in the reverse case. Through this work, we emphasize the importance of data fairness, especially if the dataset accounts for different races, with the natural requirement of clear guidelines when dealing with sensitive data, such as those related to humans.

2 RELATED WORK

Buolamwini et al. [1] showed that three commercial facial analysis systems are severely biased towards lighter-skinned subjects.

WWW '21 Companion, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8313-4/21/04.
https://doi.org/10.1145/XXXXXX.XXXXXX

Table 1: Experimental results of gender classifiers. The best accuracy is indicated in bold.

Model	Xception		MobileNet-v2		Fine-tuned Xception		Fine-tuned MobileNet-v2	
train \ test	$D^{test}_{Western}$	D^{test}_{Asian}	$D^{test}_{Western}$	D^{test}_{Asian}	$D^{test}_{Western}$	D^{test}_{Asian}	$D^{test}_{Western}$	D^{test}_{Asian}
$C_{Western}$ trained w/ $D^{train}_{Western}$	**75.38**%	72.31%	**66.15**%	63.08%	**93.85**%	73.85%	**96.92**%	76.92%
C_{Asian} trained w/ D^{train}_{Asian}	64.62%	**87.69**%	58.46%	**78.46**%	84.62%	**95.38**%	84.62%	**93.85**%

Upon evaluation of the systems on data representing white men and women, and black men and women, the authors revealed a clear bias of results towards skin color as well as gender, with a maximum error rate of 0.8% for white men, while that of 34.7% for black women. Merler et al. [10] looked into the composition of popular facial dataset [6, 9] in terms of skin color, gender and age. The authors implemented image coding techniques reflecting the diversity of facial images used to train facial recognition systems and released the Diversity in Faces (DiF) dataset with annotations. However, despite their efforts, their research lacks consideration for Asians. Moreover, several reports [5, 7] indicated that data imbalance can cause many social problems for facial recognition applications, artificial intelligence (AI)-based hiring, and criminal profiling. In this work, we extend the previous work by exploring data fairness regarding different races in particular.

Figure 1: Samples from D_{Asian} vs. $D_{Western}$, each consisting of 327 images representing female celebrities and 268 images representing male celebrities.

3 EXPERIMENTS AND RESULTS

Dataset. We crawl at least 4 images per each of 600 Asian and 600 Western celebrities from the Internet, gathering about 6,500 raw images in total. We also use real face images from FaceForensics++ [11] to augment our dataset. Then, we use Multi-task Cascaded Convolutional Networks (MTCNN) [13] to crop the face regions from the images. Some of images that are duplicated or not cropped by MTCNN are removed. Our dataset comprises 595 images representing Western people, of which 327 represent women and 268 represent men, denoted as $D_{Western}$; the same proportion holds for the dataset consisting of images representing Asian people, denoted as D_{Asian}. We split our data into train, validation, and test sets of ratio 8:1:1, ensuring no overlap of celebrities between each of the three sets. We do not apply any additional data augmentation techniques, such as horizontal or vertical flip, to improve the classification performance, as the focus of our work is to investigate a model's bias with respect to the data composition. Lastly, we also ensure that there are equally many images representing men as those representing women in the datasets. This allows us to evaluate the discrepancy in detection performance due to the racial bias in data.

Experimental setup. We train our gender classifier $C_{Western}$ with $D^{train}_{Western}$ and test separately on D^{test}_{Asian} and $D^{test}_{Western}$ (Table 1). We carry out the experiments in the same way for C_{Asian}, which is trained with D^{train}_{Asian}. We use Xception[2] and MobileNet-v2[12], since the former has shown to perform well on face recognition tasks and the latter has a light-weight use case for many practical face recognition tasks. We also suggest that fine-tuned networks from other pre-trained networks, which are widely used in practice, also produce possibly biased results under the same experimental setting as described above. In that regards, we use the pre-trained weights of ImageNet [3] to fine-tune them for our gender classification task.

We set the batch size to 64, the number of epochs to 1,000 for the non-pre-trained models and 100 for fine-tuned networks, the learning rate to 0.0001, and use the Adam optimizer [8]. We preprocess the input images by resizing them to 256×256 using bi-linear interpolation, which is commonly used for processing face images.

Preliminary results. We calculate the accuracy to assess the gender classification performance. As shown in Table 1, Xception-based $C_{Western}$ achieves 75.38% on $D^{test}_{Western}$ and 72.31% on D^{test}_{Asia}. Fine-tuned Xception-based $C_{Western}$ achieves 93.85% vs. 73.85% on $D^{test}_{Western}$ and D^{test}_{Asia}, respectively. Similar trends are observed for MobileNet-v2 with $D^{test}_{Western}$ and D^{test}_{Asia} results.

On the other hand, opposite trends are observed for Xception-based C_{Asian}, achieving 87.69% and 64.62% on D^{test}_{Asia} and $D^{test}_{Western}$, respectively. Note that the difference in classification accuracy, i.e., the bias, is larger for C_{Asian} than for $C_{Western}$. Overall, $C_{Western}$ seems less biased than C_{Asian}, but that is not the case for fine-tuned models. Also, C_{Asian} performs very poorly on $D^{test}_{Western}$. Based on these results, we can clearly see the importance and impact of data fairness in terms of composition.

4 DISCUSSION, CONCLUSION AND FUTURE WORK

In this work, we performed experiments to explore the effect of racial bias in the dataset on the classification accuracy of simple gender classifiers. The preliminary results show significant variance in accuracy, depending on the data composition regarding different races. We demonstrate that attributes, such as race and skin color, has a clear impact on the final classification performance. Therefore,

we conclude that the use of unbiased datasets, especially when face-related models are involved, is of paramount importance for robust, real-world applications. For future work, we plan to investigate the facial features that may contribute to producing biased results, such that we could make unbiased dataset or build more robust classifiers in the presence of imbalanced datasets.

ACKNOWLEDGEMENT

This work was supported by Institute for Information communication Technology Planning Evaluation (IITP) grant funded by the Korea government Ministry of Science, ICT (MSIT) (No. 2019-0-01343, Regional strategic industry convergence security core talent training business)

REFERENCES

[1] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. 77–91.

[2] François Chollet. 2017. Xception: Deep Learning with Depthwise Separable Convolutions. arXiv:1610.02357 [cs.CV]

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[4] GM Foody, MB McCulloch, and WB Yates. 1995. The effect of training set size and composition on artificial neural network classification. *International Journal of Remote Sensing* 16, 9 (1995), 1707–1723.

[5] Karen Hao. 2020. The two-year fight to stop Amazon from selling face recognition to the police. https://www.technologyreview.com/2020/06/12/1003482/amazon-stopped-selling-police-face-recognition-fight/

[6] Gary B Huang and Erik Learned-Miller. 2014. Labeled faces in the wild: Updates and new reporting procedures. *Dept. Comput. Sci., Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep* (2014), 14–003.

[7] Mit Ide. 2020. Bias In, Bias Out: Seeking More Objective Al Algorithms. https://medium.com/mit-initiative-on-the-digital-economy/bias-in-bias-out-seeking-more-objective-al-algorithms-311e1099350e

[8] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[9] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

[10] Michele Merler, Nalini Ratha, Rogerio S Feris, and John R Smith. 2019. Diversity in faces. *arXiv preprint arXiv:1901.10436* (2019).

[11] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. arXiv:1901.08971 [cs.CV]

[12] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2019. MobileNetV2: Inverted Residuals and Linear Bottlenecks. arXiv:1801.04381 [cs.CV]

[13] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters* 23, 10 (Oct 2016), 1499–1503. https://doi.org/10.1109/lsp.2016.2603342