Rufaida Khan, Julissa Mota, Gazi Rahman
Professor Zimmeck
COMP334: Information Security
4 December 2024

Title:

Bias in Facial Recognition Technology: A Case Study on FaceNet

Abstract:

The FaceNet model serves as the main subject of this paper's examination of facial recognition technology (FRT) and its associated ethical and privacy concerns, particularly within law enforcement. The study uses the Chicago Face Database to evaluate FaceNet's performance based on race and sex as documented in the accompanying Jupyter Notebook. The findings reveal significant discrepancies in accuracy, with higher false positive rates observed among Black and female individuals. The study concludes that facial recognition technology should not be used in law enforcement until its biases and potential negative effects are properly assessed and addressed.

Introduction

Facial recognition technology (FRT) consists of complex machine learning models capable of taking two images and determining whether they show the same person. This software requires the use of people's physiological data to construct these models. With FRT seeing increasing integration into law enforcement, this demands discussion. The police department of Moore, Oklahoma has already pushed out the use of cameras that collect images of people's faces as part of investigations powered by facial recognition.[1] Individuals have a right to confirm that their private data is being used responsibly as a factor in their decision on whether or not to support such policies. This paper uses FaceNet, an open-source model chosen for its fame and high accuracy, as a case study to examine FRT for biases on the grounds of race and sex per the investigation shown in Jupyter Notebook. It concludes that individuals should reject government policies that use their face data on the grounds that the current state of FRT would disproportionately harm individuals who are Black and/or Female when used in law enforcement.

FaceNet Model Overview

FaceNet is an open-source, highly performant face recognition model developed by Google researchers in 2015 to represent facial images as numerical vectors. It encodes facial features such as the eyes, nose, and lips, into mathematical representations known as

[1]Mario Gonzalez, *Moore PD technology provides accurate way for citations*, Oklahoma's New Channel 4, April 3, 2024, https://kfor.com/news/local/moore-pd-adopts-new-way-of-giving-traffi c-citations/.

embeddings.[2] For any two face images, one can determine whether they depict the same person by generating embeddings, calculating the Euclidean distance between them, and checking that the distance is low enough to confirm a match. FaceNet has become popular for facial comparison and recognition because of its state-of-the-art accuracy of 99.63% on its training data. However, the inner workings of the model are opaque; while each element of the embedding encodes some feature, researchers cannot interpret what exactly it is and thus cannot formalize the behavior of the models. This instantiates the need to conduct deeper investigations into model performance to catch unexpected behaviors early on, ones that are not captured by general metrics like overall accuracy and precision.

Data and Methodology

This study leverages the Chicago Face Database (CFD) to test differences in FaceNet's performance based on the race and sex of those depicted in an image. The CFD is a collection of 1207 images taken of 597 participants; each image is labeled with the participant's self-identified race and sex.[3] The races of the participants include Asian, Black, Latino, and White, while the sexes include Female and Male. Due to the small sample size of Asian and Latino participants, only Black and White participants were included in this study. Thus, the constrained dataset consisted of 990 images taken of 380 participants; 104 Black Females, 93 Black Males, 90 White Females, and 93 White Males. This provides focus into the examination, comparing how FaceNet might perform among the listed demographic groups.

The study consists of a pairwise comparison of every possible image pair in the dataset. The model was used to generate embeddings of every image. Then, for every embedding, the distance was calculated between it and every other embedding. For each pair, the prediction was 'True' if the distance between their embeddings was less than FaceNet's conventional threshold of 1, suggesting the images depicted the same person, and false otherwise. For clarity, two images will be referred to as a 'match' if they show the same person; otherwise, they will be referred to as 'not a match'. The model predictions were categorized as follows: a 'true positive' when the model correctly predicted a match, a 'false positive' when the model incorrectly predicted a non-match as being a match, a 'true negative' when the model correctly predicted a non-match, and a 'false negative' when it incorrectly predicted a match as being a non-match. The 'false positive rate' marks the rate at which the model predicts true for image pairs that are not actually matches; it is taken by dividing the number of mislabeled non-matches by the total number of non-matching pairs. The 'false negative rate' marks the rate at which the model predicts false for image pairs that are actually matches; it is taken by dividing the number of images mislabeled matches by the total number of matching pairs. At each phase of analysis, FaceNet was assessed based on its accuracy, distribution of its predictions among each of the listed categories, and rate of false positives. An analysis of the rate of false negatives was omitted due to the lack of false negatives experienced by FaceNet. First, the model's

---

[2]Florian Schroff, Dmitry Kalenichenko, and James Philbin, *FaceNet: A Unified Embedding for Face Recognition and Clustering*, Cornell University, March 12, 2015, https://arxiv.org/abs/1503.03832?utm_source=chatgpt.com.

[3]Debbie S. Ma, Joshua Cornell, and Bernd Wittenbrink, The *Chicago face database: A free stimulus set of faces and norming data*, Wittenbrink, January 13, 2015, https://www.wittenbrink.org/cfd/mcw2015.pdf.

performance was examined cumulatively across every image pair. Then, FaceNet's performance was examined based on race, constraining the results to only include image pairs where both images depicted Black people and then those where both depicted White people. Afterward, the model's performance was examined based on sex, constraining results to include image pairs where both images depicted Females and pairs that depicted Males. Finally, the model's performance was examined based on both race and sex, constraining results to include image pairs where both images depicted Black Females, Black Males, White Females, and White Males.

Results

The cumulative performance of FaceNet was comparable to its performance upon release. This study employed a version of the model trained on the VGGFace2 dataset, performing with 99.65% accuracy during its original training.[4] During the study, FaceNet performed with a similar accuracy of 97.85%. The small dip in performance can be attributed to a predictable dip when models are repurposed into real-world applications, thus not warranting concern. Out of every image pair, FaceNet impressively only experienced a single false negative. That is to say, the model was excellent at determining true positives, at correctly predicting true when two images were a match. In contrast, the model dipped in performance when determining true negatives, when accurately predicting false for images that were not a match. Overall, the model experienced a false negative rate of only 0.07%, compared to a false positive rate of 2.16%. While this may seem promising at face value, FaceNet's efficacy is discredited by the fact that the model proceeded to show different metrics based on demographic information.

Race

FaceNet demonstrated statistically significant discrepancies in performance between image pairs that depicted Black people compared to those that depicted White people. It exhibited an accuracy of 92.73% for Black image subjects compared to 99.59% for White subjects. The single false negative mentioned earlier occurred in the group of White individuals, leading to a 0% rate of false negatives for Black people compared to a rate of 0.15% for White people. This is not a significant metric given that there was a single false negative in the entire study. However, more remarkably, FaceNet exhibited far more false positives for Black people than White people. The model had a 7.32% rate of false positives for the former compared to 0.41% for the latter. This marks a discrepancy of 6.91%. In the context of law enforcement, FaceNet would suggest the arrest of the wrong person 6.91% more frequently for Black suspects compared to White counterparts. Performing a Chi-Square test on this discrepancy yielded a Chi-Square value of 6962.20 and a p-value of 0.0, confirming with confidence that FaceNet identified false positives at a statistically significant higher rate for Black image subjects than for White subjects.

---

[4]Tim Esler, et al., *Pretrained Pytorch face detection (MTCNN) and facial recognition (InceptionResnet) models*, GitHub, August 2, 2024, https://github.com/timesler/facenet-pytorch/activity.

Sex

Facenet exhibited statistically significant discrepancies in performance between image pairs involving Females compared to those involving Males. Although the differences were less pronounced, the model exhibited 94.36% accuracy in correctly predicting the relationship between images of Females compared to a 98.1% accuracy for Males. Among its guesses, the model experienced a 0% rate of false negatives for Females and 0.15% for Males. The model also experienced a false positive rate of 5.68% for Females compared to 1.91% for Males, marking a 3.77% difference. As with the racial study, this means that FaceNet would suggest the arrest of the wrong person 3.77% more often for Female suspects than Male suspects. This discrepancy in false positives was confirmed to be statistically significant via a Chi-Square test, yielding a Chi-Square value of 2174.33 and p-value of 0.0. The decrease in the Chi-Square value compared to the racial section reflects the less pronounced yet still significant variation in performance.

Race and Sex

Facenet showed the greatest statistically significant discrepancies in performance when tested based on the intersection between race and sex. When image pairs were constricted to those pairs that both depicted Black Females, Black Males, White Females, and White Males respectively, the model demonstrated the greatest rate of inaccuracy towards those who were Black Females. The results are as follows: Black Females experienced an accuracy of 82.23% and false positive rate of 17.97%, Black Males had an accuracy of 93.33% and false positive rate of 6.75%, White Females experienced an accuracy of 99.24% and false positive rate of 0.77%, and White Males had an accuracy of 99.16% and false positive rate of 0.85%. These results demonstrate that White image subjects are generally accurately judged by FaceNet regardless of sex. However, Black image subjects experienced more inaccuracy and higher false positives, which were worsened for those who were both Black and Female. In the context of law enforcement, this means that the intersection between race and sex would be a strong factor in whether FaceNet would suggest the arrest of the wrong individual.

Discussion

Implementing the use of FaceNet in law enforcement without addressing its evident biases would be unethical, as it would enact race- and sex-based harm. This study has demonstrated that FaceNet is more likely to predict false positives when identifying Black and Female individuals. False arrests resulting from such errors have severe consequences, including damage to one's livelihood, reputation, and health.[5] While one might argue that false arrests are an unfortunate yet natural part of the legal process, the consequences should at the very least be uniformly distributed among members of a society. At an individual level, compromising just one person for the benefit of everyone is unfair. In this case, compromising groups of people based on demographics would be a blanket undervaluation of the rights of people on the grounds of race and sex, qualifying as racism and sexism.

---

[5]Kathryn Campbell and Myriam Denov, *Miscarriages of Justice: The Impact of Wrongful Imprisonment*, Government of Canada, January 20, 2023, https://www.justice.gc.ca/eng/rp-pr/jr/jr13/p5a.html.

FaceNet's performance across demographic groups offers a case study into the dubious ethics of using FRT in law enforcement. Despite its acclaim and accuracy, it demonstrates that even highly performant models require investigation into their inner workings. While flaws in any facial recognition model may be attributed to the data it is trained on, it is up to the agencies that seek to use them to prove that their instance of FRT would be unbiased and would impact all members of a community uniformly. Only then can a fair discussion take place on whether or not to implement these models. Without this evidence being provided, the risk of compromising community members on the grounds of race and sex makes it such that these models should not be formally adopted.

Conclusion

This case study serves as a manifesto to demonstrate the potential consequences of sharing one's physiological data with government agencies for the purpose of implementing FRT; in offering this information, it serves to help readers make an informed decision on whether to accept such practices in their own communities. The demonstration that FRTs can corroborate race- and sex-based harm proves to readers that they should not accept such policies without demanding further context on the penchant for algorithmic bias in the technology being implemented. This should offer important context when making decisions about one's data and choosing whether to disclose their information in the future. The study certainly has room for improvement. As discussed in the accompanying Jupyter Notebook, it can benefit from a larger sample size of images, more races being incorporated, and certainly more models being implemented. It is important to recognize that FaceNet is not representative of all models; it represents simply one instance of FRT. This investigation into FaceNet simply proves that FRT engineers have a responsibility to do further research into the potential consequences of their models and prove that their software will behave responsibly.

Bibliography

Campbell, Kathryn and Denov, Myriam. *Miscarriages of Justice: The Impact of Wrongful Imprisonment*. Government of Canada. January 20, 2023. https://www.justice.gc.ca/eng/rp-pr/jr/jr13/p5a.html.

Esler, Tim, et al. *Pretrained Pytorch face detection (MTCNN) and facial recognition (InceptionResnet) models*. GitHub. August 2, 2024. https://github.com/timesler/facenet-pytorch/activity.

Gonzalez, Mario. *Moore PD technology provides accurate way for citations*. Oklahoma's New Channel 4. April 3, 2024. https://kfor.com/news/local/moore-pd-adopts-new-way-of-giving-traffic-citations/.

Ma, Debbie S., Joshua Cornell, and Bernd Wittenbrink. The *Chicago face database: A free stimulus set of faces and norming data*. Wittenbrink. January 13, 2015. https://www.wittenbrink.org/cfd/mcw2015.pdf.

Schroff, Florian, Dmitry Kalenichenko, and James Philbin. *FaceNet: A Unified Embedding for Face Recognition and Clustering*. Cornell University. March 12, 2015. https://arxiv.org/abs/1503.03832?utm_source=chatgpt.com.