Investigating Accuracy Disparities for Gender Classification Using Convolutional Neural Networks

Lia Chin-Purcell
School of Information
UC Berkeley
Berkeley, USA
lchinpurcell@ischool.berkeley.edu

America Chambers
Department of Computer Science
University of Puget Sound
Tacoma, WA
alchambers@pugetsound.edu

***Abstract*—Automatic gender recognition (AGR) is a subfield of facial recognition that has recently been scrutinized for bias in the form of misgendering and erasure against various identity groups in our society. Recent studies have found that several commercial AGR classifiers (from Microsoft, IMB, Face++) are biased against women and darker-skinned people as well as gender non-binary people [8, 11]. In this work, we investigate and quantify AGR classifier bias against transgender people by developing and evaluating three different convolutional neural networks (CNN): using images of cisgender individuals, using images of transgender individuals, and using images of both cisgender and transgender individuals. We find that the cisgender trained classifier is 91.7% accurate when evaluated on cisgender people, but only 68.9% accurate when evaluated on transgender people, with the worst performance of 38.6% precision for transgender men. We investigate this low precision further by performing additional experiments where various parts of the face are obscured. We end with recommendations for commercial classifiers based upon our findings.**

***Index Terms*—facial recognition, transgender, convolutional neural network**

I. Introduction

Research has shown that data-driven artificial intelligence systems often reflect the biases of a given society [1], [2], [4], [9], [14]. An example of this phenomenon is facial classification systems which are trained on images of human faces to detect attributes such as emotion, gender, criminality, and race [6], [9]. Automatic gender recognition (AGR), a type of facial classification, attempts to predict an individual's gender from an image or video. In [4], the authors found that commercial AGR systems from Microsoft, IBM, and Face++ perform "best for lighter individuals and males overall" and perform "worst for darker females". In addition, because of the difficulty in obtaining training data, many AGR classifiers are built to predict a binary gender output and are trained on mostly cisgender[1] white males [11]. As a result, AGR classifiers frequently misgender transgender individuals [11].

This phenomenon is concerning in light of the increasingly central role that facial recognition systems play in our daily lives. Today, facial recognition is used as a tool for mass surveillance in airport security [2], by local law enforcement agencies [3], as a way to unlock a smartphone, in social media analytics [4], and more. Given the increasing role of facial recognition systems, the higher rates of misclassification for minority individuals (e.g. transgender individuals) has the potential to perpetuate discrimination [11].

Recent research on this topic has focused on creating debiasing algorithms [13], [14] and quantifying the amount of bias present in a given classifier [4]. In a similar vein to such work, we investigate the degree to which the images used to train an AGR classifier impact the misclassification of cisgender versus transgender individuals. In particular, we train three different convolutional neural networks (CNN): using images of cisgender individuals, using images of transgender individuals, and using images of both cisgender and transgender individuals. Each classifier is then evaluated by calculating the accuracy, precision, and recall on *all* three datasets. By employing transfer learning – i.e., training on one dataset of images and testing on a different dataset of images – we mimic the real-world usage of AGR systems which are often trained on images of mostly cisgender individuals and then deployed on the general public.

Interestingly, we find that transgender women face the highest rates of misclassification and we investigate this phenomenon further by obscuring various parts of the face in order to better understand the trained CNNs. We end the paper with recommendations for commercial classifiers in order to mitigate the effects caused by classification algorithms trained on imbalanced datasets.

II. Datasets

We compiled two different facial image datasets: one consisting of cisgender individuals, and the other consisting of transgender individuals. In light of recent critiques [16] of papers such as [10] which use images taken from transgender

[1]Throughout this paper, we adopt the terminology of Ahmed et al. [1] We refer to an individual's gender as their self-identification as a "man, woman, or anywhere outside that binary" [1]. We use the term *transgender* or *trans* to "describe individuals whose gender does not conform to expectations surrounding the one assigned to them at birth" [1]. We use the term *cisgender* to describe individuals whose gender does correspond to the one assigned to them at birth.

978-0-7381-1423-1/21/$31.00 © 2021 IEEEE

Fig. 1. Example Images from Datasets

youtubers without permission, we only collected images of public figures – e.g., actresses, actors, singers, politicians.

A. Cisgender Dataset

The cisgender dataset was collected from images posted on IMDb [8] that had accompanying age and gender information. Example images from the dataset are shown in Figure 1 on the bottom row. The dataset includes individuals who are aged 18 and older. The training set contains 632 images of men and 818 images of women. The test set contains 181 images of men and 274 images of women (a 75/25 split into training and testing). There were not exact estimates on race from the source website, but from a diversity report on Hollywood [5] we estimate that approximately 78% of the cisgenderdataset is white and 22% are individuals of color (non-white).

B. Transgender Dataset

The transgender dataset consists of public figures who identify either as transgender men or transgender women. Example images from this dataset are shown in the top row of Figure 1. It was difficult to construct a sufficiently large dataset since we only used images of public figures. As such, we collected multiple images of the same individual to increase the size of the transgender dataset. We collected images from 19 men and 24 women (approximately 35 images per individual) scraped from Instagram. The dataset includes individuals aged 18 and older with the majority of individuals in the 20-35 age range. The training set contains 586 images of men and 843 images of women. The test set contains 216 images of men and 274 images of women (again, a 75/25 split into training and testing). We intentionally compiled a racially diverse set of images. The transgender dataset consists of 21% white individuals and 79% individuals of color. Table I shows a summary of both datasets.

TABLE I
DATASET DIVERSITY.

Cisgender dataset		
	Men	Women
Images	813	1092
Percent non-white	24%	28%
Transgender dataset		
	Men	Women
Images	801	1117
Percent non-white	60%	92%

III. Convolutional Neural Network Classification Model

A convolutional neural network (CNN) is a type of neural network that is designed to work specifically with image data (allowing for a reduction in the number of parameters that must be learned from training data). CNNs are composed of a number of "layers" where the first layer is known as the "input layer", the last layer is known as the "output layer" and the intermediate layers are known as the "hidden layers".

The input layer is a 3-dimensional tensor where the dimensions correspond to the height of the image, the width of the image, and the number of color channels (e.g. a color image represented using an RGB color model would have 3 color channels). The output layer encodes a probability distribution over the possible labels of an image – in this case, *man* or *woman*.

The hidden layers represent transformations of 3-dimensional tensors. That is, each hidden layer accepts a 3-dimensional tensor from the previous layer, applies a mathematical function, and produces as output a transformed 3-dimensional tensor. The standard layer types are convolutional layers, pooling layers, and fully-connected layers. These layers can be combined in various orderings to produce different CNN architectures. In this way, the architecture of a CNN is analogous to creating a stack of Lego pieces where each blue piece represents a convolutional layer, each red represents a pooling layer, etc.

In a convolutional layer, a small matrix – known as a *filter* – is repeatedly applied to the input tensor from left to right and top to bottom. Figure 2 shows an example 3x3 filter being passed over a 2-dimensional image. If we interpret the numbers in the image to be grayscale values, then the purpose of the filter is to detect vertical edges in the image.

In the first iteration (leftmost), we apply the filter to the top-left of the image. An element wise multiplication is performed and the values are summed to produce a scalar output – in this case, the output is $9(1) + 9(1) + 7(1) = 25$. In the second iteration (middle), the filter is moved over by 1 pixel and applied again. The output of the second iteration is 0. In the next iteration (rightmost), the filter is moved down by 1 pixel and reset to the left side of the image. The output in this case is $9(1) + 7(1) + 8(1) = 24$. In the final iteration (not shown), the filter is moved over once more by 1 pixel. In this way, the filter visits overlapping regions of the input image. The final output of the convolution is then given by the 2x2 matrix:

$$\begin{bmatrix} 25 & 0 \\ 24 & 0 \end{bmatrix}$$

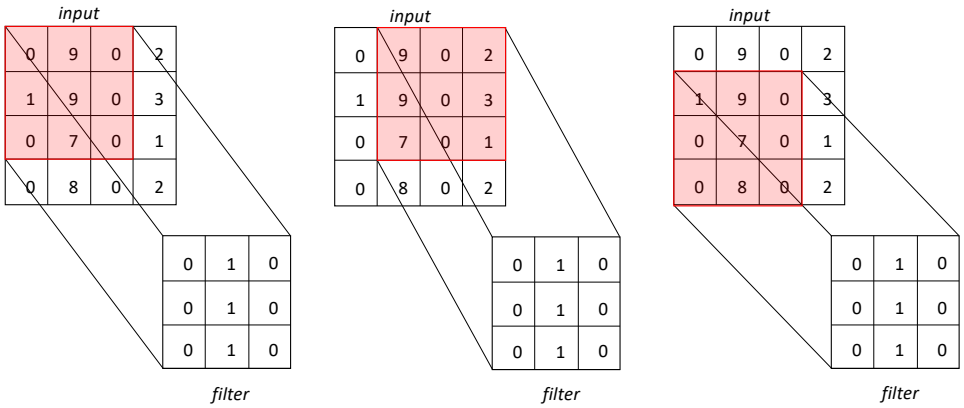

Fig. 2. The first three iterations of a 3x3 filter being passed over a 2-dimensional image

Note that the values of the filter – i.e. the numbers in the matrix – are learned during training time.

A pooling layer is meant to reduce the size of the resulting 3-dimensional tensor. In a pooling layer, we again sweep over the input image however we do not visit overlapping regions of the image. In each iteration, a reducing operation is applied to a subregion of the image. Common reducing operations include averaging (i.e. averaging all values in the subregion), or the max function (i.e. taking the max of all values in the subregion). Figure 3 shows an example of a 2x2 max pooling layer being applied to a 2-dimensional image. The final output of the max pooling layer is given by the 2x2 matrix:

$$\begin{bmatrix} 9 & 3 \\ 8 & 2 \end{bmatrix}$$

Pooling provides a form of summarization and reduces the size of the output tensor (thus reducing the number of weights of the neural network).

Finally, a fully-connected layer is the same as a fully-connected layer in a feedforward neural network where the nodes of one layer are fully connected to the nodes in the next layer. Fully-connected layers are usually used as the last layers in a CNN when we desire to transition from 3-dimensional tensors to a 1-dimensional vector representing a probability distribution over the labels.

A. Pre-Training

Pre-training a CNN refers to the practice of first learning the weights and parameters of the CNN using an extremely large database of images related to, but distinct from, the target task. This is especially useful when the dataset of images for the target task is smaller – e.g., our combined cisgender and transgender datasets contain only $3,823$ images. In contrast, ImageNet [18], a common dataset that is used for pre-training CNNs, contains over 1 million images.

The AGR classifier models used in this paper were built using the fast.ai framework [7] which sits on top of Pytorch [12] – a Python library for deep learning. The fast.ai library provides programmers with two different pre-trained CNN models which serve as a foundation for our AGR classifiers. During training, we update the final layers of the pre-trained CNN using the images and labels in our data sets.

We used the resnet-34 model [19] with one-cycle learning, proven to be a very effective learning method [7]. We train 3 different versions of the resnet-34 classifier: one using the cisgender dataset, one using the transgender dataset, and one using both datasets. We used square cropping of the images, no image transformations, and a batch size of 64. All parameters were held constant across the 3 versions.

The resnet-34 model was pre-trained on the ImageNet data set [18]. The images in ImageNet were collected on the internet and are thus subject to existing bias, however, this training set is very broad and not limited to gender classification. Typically, this pre-training allows the model to pick out features in an image such as edges and shapes, while our training of the final layers is specific to gender classification.

Finally, note that the input for the classifier is an image of a face, and the output is a vector $[p(man), p(woman)]$ where $p(man)$ is the probability that the image is a man and $p(woman)$ is the probability that the image is a woman. These values sum to 1.

IV. Results

For each classifier, we predict the gender of the images in each test set. For example, the CNN trained on the cisgender data set is then used to predict the gender of the images in the cisgender, transgender, and combined test sets. In this way, we are performing a type of transfer-learning and are mimicking the real-world usage of AGR systems that are often trained primarily on cisgendered individuals and then deployed on the general public.

Tables II through VI show the accuracy, precision, and recall for each of the 3 classifiers. The accuracy is the total percent of images in the test set that are correctly labeled. The precision for the label *man* is the number of images correctly assigned

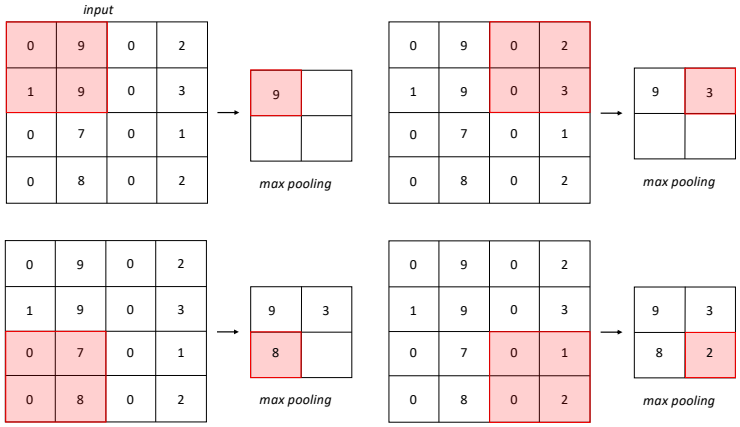

Fig. 3. A 2x2 max pooling layer being applied to a 2-dimensional image

the label *man* divided by the total number of images (correctly or incorrectly) assigned the label *man*. The recall for the label *man* is the number of images correctly assigned the label *man* divided by the total number of images whose ground truth label was indeed *man*. Precision and recall for the label *woman* is defined similarly.

Figure 4 provides a graphical illustration of the precision and recall for the combined and cisgender classifiers on the transgender test set specifically. The unshaded circle represents the images in the test set whose ground truth label was *man* or *woman* respectively. The shaded circle represents the images for which the classifier predicted *man* or *woman* respectively. The intersection of the two circles as a proportion of the total area of the shaded circle represents the precision of the classifier (labeled using an arrow). The intersection of the two circles as a proportion of the total area of the unshaded circle represents the recall of the classifier (labeled inside the circle).

TABLE II
OVERALL ACCURACY.

	Test set		
Training set	Cisgender	Transgender	Combined
Cisgender	91.7%	68.9%	83.5%
Transgender	75.4%	88.3%	82.1%
Combined	91.2%	86.9%	89.1%

TABLE III
PRECISION FOR MEN.

	Test set		
Training set	Cisgender	Transgender	Combined
Cisgender	89%	**38.6%**	66.5%
Transgender	74%	80.1%	77.3%
Combined	85.3%	76.9%	79.8%

V. ANALYSIS OF RESULTS

Not surprisingly, the cisgender and transgender classifiers were most accurate when evaluated on their respective test sets – with accuracies of 91.7% and 88.3% respectively.

TABLE IV
RECALL FOR MEN.

	Test set		
Training set	Cisgender	Transgender	Combined
Cisgender	89.9%	72.9%	92%
Transgender	67.3%	92.5%	79.5%
Combined	91.7%	93%	93.2%

TABLE V
PRECISION FOR WOMEN.

	Test set		
Training set	Cisgender	Transgender	Combined
Cisgender	93.4%	90%	95.8%
Transgender	76.3%	94.9%	85.6%
Combined	94.9%	93.5%	95.8%

TABLE VI
RECALL FOR WOMEN.

	Test set		
Training set	Cisgender	Transgender	Combined
Cisgender	92.8%	67.8%	79.8%
Transgender	81.6%	85.8%	83.9%
Combined	90.9%	78.2%	86.8%

The combined classifier had the highest accuracy on the cisgender test set and performed worse on the transgender test set. In terms of precision and recall, the combined classifier obtained a precision of 76.9% and a recall of 93% for transgender men. A lower precision but higher recall indicates that the combined classifier over-predicted the label *man* and thus under-predicted the label *woman*. This is further supported by the combined classifier's high precision but lower recall for transgender women.

One result that particularly stands out is the cisgender classifier's poor performance on the transgender test set with 68.9% accuracy. Investigating further, we see that the cisgender classifier attained only 38.6% precision and 72.9% recall for transgender men (see Table III). This suggests that the phenomenon of over-predicting the label *man* exhibited by the combined classifier – which had access to images of

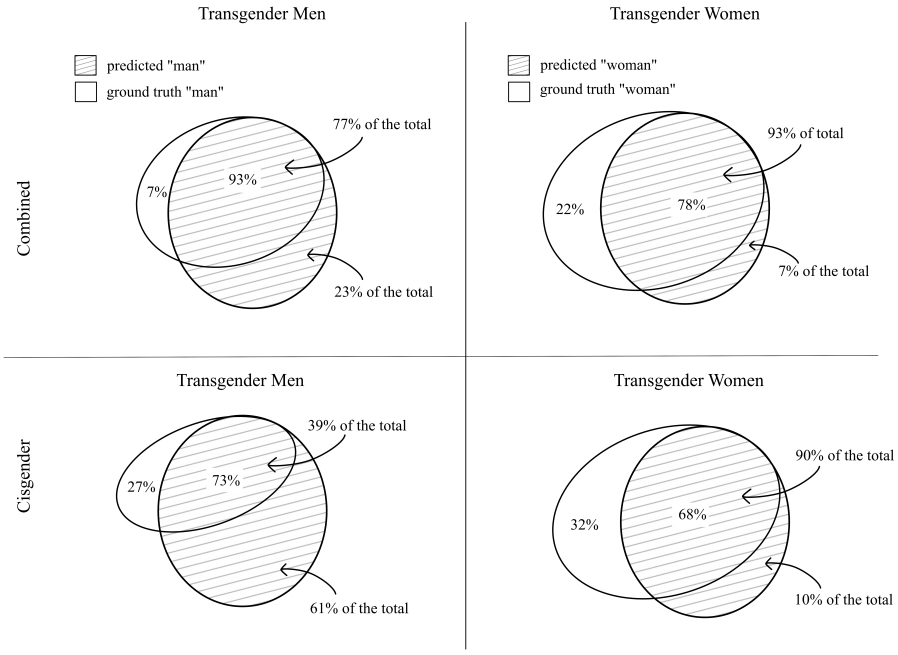

Fig. 4. An illustration of the precision and recall for the combined and cisgender classifiers on the transgender test set.

both cisgender and transgender individuals during training – is exhibited to an even greater degree by the cisgender classifier, which was trained on images of cisgender individuals alone. Overall, the results suggest that by adding images of transgender individuals to the training set, the tendency to over-predict the label *man* is reduced but not eliminated.

In the next sections, we investigate possible reasons for why the cisgender classifier performed so poorly when classifying transgender men by evaluating each classifier on images of faces where the features have been obscured, as well as evaluating the cisgender classifier on white transgender men versus transgender men of color.

VI. Visualization of the Models

Fig. 5. Example Obscured Images

To gain a deeper understanding of the differences between the cisgender and transgender classifiers, we visualized the models by obscuring either the eyes, eyebrows, nose, or mouth of each individual in the transgender *test* set, and then computed the average confidence of the classifier. The average confidence of the classifier was computed by averaging $p(man)$ for the images in the test set whose labels were indeed man, and averaging $p(woman)$ for the images in the test set whose labels were indeed woman. Similar methods of obscuring have been conducted in [15]. Table VII and Table VIII show the average confidence for the transgender and cisgender classifiers respectively.

A. Transgender Classifier

For the transgender classifier, it is clear from Table VII that the eyes are a strong predictor of the label *woman*. When obscuring the eyes, the average confidence of the transgender classifier drops drastically from 0.9858 (for unobscured images) to 0.5420. For men, there was no one feature that greatly affected the classifier performance although obscuring the eyebrows and mouth caused a slight decrease in the average confidence from 0.9976 to 0.9432 and 0.8988 respectively.

B. Cisgender Classifier

For the cisgender classifier, the overall low average confidence for each feature for the men (see the last column in Table VIII) is perhaps not surprising given the low precision noted in the previous section. For images labeled *man*, obscuring any feature other than the mouth actually *increases* the classifiers confidence in predicting the label *man*. We further expound upon each feature below.

Obscuring the eyes and eyebrows for women led to a decrease in average confidence from 0.9858 to 0.8611 and 0.8834 respectively. Correspondingly, obscuring the eyes and eyebrows for men led to an *increase* in average confidence from 0.4155 to 0.6326 and 0.5637 respectively. Taken together, this seems to imply that the presence of the eyes and eyebrows are used by the classifier as evidence supporting

the prediction of the label *woman* regardless of the ground truth. This may be due to the fact that transgender women can modify their eyes and eyebrows (e.g. with makeup) which may be a feature that the cisgender classifier uses in the prediction of *woman* while for transgender men it may be more costly to modify the physical shape of their eyes and eyebrows.

Obscuring the mouth for men led to a decrease in average confidence from 0.4155 to 0.2826 implying that the mouth is an influential feature used by the classifier as evidence for assigning the label *man*.

Interestingly, the nose is a confounding feature since when it is obscured both the average confidence for women and men increases. Again, this may be due to the fact that it is costly to modify the physical shape of a nose. For transgender individuals, then, the nose alone is not a good indicator for the classifiers prediction of gender and removing it from the image removes a source of noise.

Overall, these findings suggest that features that are cultural indicators of gender and easily modified (e.g. eyes, eyebrows, and mouths) are more robust indicators for the model's gender classification of transgender individuals than physical features that are more costly to modify (e.g noses).

TABLE VII
AVERAGE CONFIDENCE OF THE **transgender** CLASSIFIER

Obscured	Average Confidence in Predicting Woman	Average Confidence in Predicting Man
Eyes	0.5420	0.9962
Eyebrows	0.9020	0.9432
Nose	0.8883	0.9788
Mouth	0.9115	0.8988
Non-obscured	0.9858	0.9976

TABLE VIII
AVERAGE CONFIDENCE OF THE **cisgender** CLASSIFIER

Obscured	Average Confidence in Predicting Woman	Average Confidence in Predicting Man
Eyes	0.8611	0.6326
Eyebrows	0.8834	0.5637
Nose	0.9346	0.5279
Mouth	0.9198	0.2826
Non-obscured	0.9270	0.4155

VII. EVALUATION BROKEN DOWN BY SKIN TONE

TABLE IX
AVERAGE CONFIDENCE OF THE CISGENDER CLASSIFIER FOR MEN BROKEN DOWN BY SKIN TONE

	Non-White Transgender men	White Transgender men
Average probability man	0.4978	0.4584

In a similar vein to the analysis in Section VI, we investigate if the cisgender classifier's poor performance on transgender men was in part due to the imbalance of the two data sets racially (see Table I). Although a more comprehensive and specific approach would be to categorize the transgender men based on a spectrum of skin tones [4], we concluded that if skin tone was the reason the classifier was performing so poorly, a pattern would still arise from simply categorizing images as white or non-white.

Table IX shows the average confidence of the cisgender classifier for non-white transgender men and white transgender men (for the un-obscured images in the original transgender test set). Interestingly, the cisgender classifier is almost equally ambivalent in its classification of both groups with an average confidence of 0.4978 for non-white transgender men and 0.4584 for white transgender men. This suggests that skin tone was not a significant reason for the cisgender classifier's poor performance on transgender men.

VIII. CONCLUSION AND RECOMMENDATIONS

We evaluated three different convolutional neural networks trained using images of cisgender individuals, using images of transgender individuals, and using images of both cisgender and transgender individuals. Each classifier performed relatively well when evaluated on its own test set. When evaluated on the transgender test set, however, the combined and cisgender classifiers both tended to over predict the label *man*.

From our visualization of the classifiers, this shortcoming was not due to the cisgender dataset being overwhelmingly white, as the classifier did equally poorly on both white and non-white transgender men. We found that the eyes (and to a lesser extent the eyebrows) played a strong role for both the transgender and cisgender classifier in assigning the label *woman*. We also found that for the cisgender classifier, the nose was a confounding feature for both labels.

Overall, these results suggest that an AGR system trained on predominately cisgender individuals is more likely to misgender transgender individuals by over predicting the label *man* thus mislabeling transgender women. While the inclusion of transgender individuals in the training set (as seen for the combined classifier) mitigates the model's error when evaluated on transgender individuals, this alone is not enough to achieve the same level of accuracy for transgender individuals as was seen for cisgender individuals.

This discrepancy in accuracy and over-prediction of the label *man* is interestingly reflected in society as well when humans, and not algorithms, are asked to classify transgender individuals. That is, when humans are the "classifiers", transgender women are often seen as more conspicuous than transgender men – "a 6'2" woman is often more conspicuous than a 5'4" man" [17]. This highlights an important issue that developers of commercial AGR systems should also consider: AGR systems may perpetuate existing trends of misgendering of transgender people overall, and particularly misgendering transgender women.

Finally, we note that the vast majority of AGR systems perform binary classification (i.e. *man* or *woman*) which automatically excludes transgender people who do not identify within this gender binary. This includes people who identify as non-binary, gender-fluid, agender, etc.

In order to mitigate potential harm, commercial AGR classifiers trained on a majority white, cisgender, male population should conduct audits of their classifier's performance on transgender individuals, paying particular attention to the types of misclassifications that occur and the features that are being used by the classifier. In addition, adding images of transgender individuals to the training set, in and of itself, is not necessarily sufficient to eliminate the gap between prediction accuracy on cisgender and transgender individuals.

IX. Future Work

Given the results from obscuring parts of the face for the images in the transgender test dataset, it would be interesting to repeat these obscuring experiments on the images in the *cis* test dataset to help validate our conclusions. Also, when obscuring images, we used rectangular bounding boxes that potentially covered other areas of the face that may have been important to the classifier – e.g., when obscuring the nose we may also obscure facial hair just below the nose. It would be interesting then to repeat the obscuring experiments using a more finely drawn bounding box that obscured only the feature itself.

Our analysis of race and skin-tone as potential reasons for over fitting and mis-classifiation was directly addressing the cisgender model's poor performance on the transgender test set. However, future work could include a more comprehensive analysis of race and skin tone in the auditing of the classifiers. In particular, we would perform a nuanced breakdown of skin tone and examine how the proportion of people of color within the *training set* affects the performance of the classifier.

Finally, it was difficult to construct a sufficiently large dataset of transgender individuals since we only used images of public figures. As such, we had to include multiple images of the same individual in the transgender dataset. This surely had an impact on our results and, if possible, it would be interesting to re-train the transgender classifier using a dataset of more unique individuals.

Acknowledgment

References

[1] A. Ahmed, "Trans Competent Interaction Design: A Qualitative Study on Voice, Identity, and Technology, Interacting with Computers," vol. 30, Issue 1, pp. 53–71, 2018.

[2] A. Khalid, "Facial Recognition May Boost Airport Security But Raises Privacy Worries," NPR, 2017.

[3] D. Harwell, "Oregon Became A Testing Ground For Amazon's Facial-Recognition Policing. But What If Recognition Gets It Wrong?," The Washington Post, 2019.

[4] J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," Prof. of Machine Learning Research, 2018.

[5] D. Hunt, A. Ramon, M. Tran, A. Sargent, and D. Roychoudhury, "Hollywood Diversity Report", UCLA College of Social Sciences, 2018.

[6] Face++ API. https://www.faceplusplus.com/. Accessed 2019-8-04.

[7] fast.ai API. https://www.fast.ai/. Accessed 2019-8-04.

[8] IMDB-WIKI dataset. https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/. Accessed 2019-8-04.

[9] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias: There's software used across the country to predict future criminals. And its biased against blacks," ProPublica, 2016.

[10] V. Kumar, R. Raghavendra, A. Namboodiri and C. Busch, "Robust transgender face recognition: Approach based on appearance and therapy factors," pp. 1–7, 2016.

[11] O. Keyes, "The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition," Proc. ACM Hum.-Comput. Interact, 2018.

[12] Pytorch API. https://pytorch.org/. Accessed 2019-8-04.

[13] A. Amini, A. Soleimany, W. Schwarting, S. N. Bhatia, and D. Rus, "Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure," Proc. of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19). Association for Computing Machinery, pp. 289–295, 2019.

[14] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," In Advances in Neural Inf. Processing Systems, 29, pp. 4349—4357, 2016.

[15] M.D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," In European Conf. on Computer Vision, 2014.

[16] J. Vincent, "Transgender YouTubers had their videos grabbed to train facial recognition software," The Verge, 2017.

[17] C. Alter, "Seeing Sexism from Both Sides: What Trans Men Experience," Time, 2016.

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," IEEE Conf. on Computer Vision and Pattern Recognition, pp. 248–255, 2009.

[19] K. He, X. Zhang, S. Ren, and J Sun. "Deep residual learning for image recognition," IEEE Conf. on Computer Vision and Pattern Recognition, pp. 770–778, 2016.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in Neural. Info. Processing Systems, vol. 25, pp. 1097–1105, 2012.