Gender Classification by Race using Convolutional Neural Networks

Nicholas Perello
University of Massachusetts Amherst
140 Governors Dr., Amherst, MA 01003
nperello@umass.edu

1. Introduction

As automated programs utilizing neural networks have increased over the past several years, there has been a coinciding increase in concern that such programs will produce outputs that are unintendedly biased towards individuals. These biases may include excluding individuals with certain names from loan applications, over-predicting that a people of one race will have higher chances to recommit crimes, and scoring applicants from historically women colleges as worse candidates. Outcomes like this are not what the developers of automated programs intended and often stem from uncaught caveats from training datasets. Many datasets may not factor for unique features of individuals such as race, gender, and socioeconomic status, and how these features have important contexts that a machine would not intrinsically know. With the ever-increasing popularity and everyday reliance of automated facial recognition software in people's everyday lives, it is imperative that facial recognition software does its best to avoid unintended biases.

In 2018, J. Buolamwini & T. Gerbu [2] found that for three commercial facial recognition and gender classification systems, gender misclassification occurred at higher rates for all three when predicting for darker-skinned females. This higher misclassification rate occurred even though these systems used a dataset of people of all races and genders. While the three companies responsible for these systems have done work to close this misclassification rate, this project aims to explore a new way to prevent such biases. This project's proposed approach for the mitigation of bias in gender classification based on skin-color is to create two convolutional neural networks that work sequentially. The first will predict the race the inputted individual's image while the second will predict the gender of the individual based on training with faces of individuals of a similar race. The inspiration for this approach was based on the idea that historically underrepresented racial groups would tend to have different and perhaps worse data than majority racial groups which could then affect training neural network models in many differing applications. It may be possible that comparing data within these underrepresented groups could produce more fair results than comparing it with the majority groups. In the context of facial recognition and gender classification, this would be similar to comparing people of different races who have different facial characteristics. By comparing individuals with others of similar race and facial structure, ideally, better accuracies in gender classification may be achieved. It is hoped that with this project, an inspiration to study methods similar to the proposed method is achieved for other applications of neural networks such that bias in essential automated software can be prevented.

2. Problem Statement

The goal of this project is to improve the gender classification of people, especially for those of races that are of darker skin colors. To achieve this, datasets that included faces with labeled races/ethnicities and gender was needed. Unfortunately, the Gender Shades dataset from Buolamwini & Gerbu. [2] was unavailable and alternatives were required. Therefore, datasets that are to be used is UTKFace [4] and 10k U.S. Adult Faces Database [1]. The UTKFace dataset consists of over 20,000 images of faces of individuals of all ages that are labeled by age, gender, race/ethnicity, and date the image was taken. For this dataset's race label, there are 5 categories represented as integers from 0 to 4 denoting White, Black, Asian, Indian, and Others (like Hispanic, Latino, Middle Eastern). For the purpose of this project, the pre-aligned and cropped faces set of images from this dataset will be used. These are the same images of faces but have been cropped to 200x200 RGB images aligned to the face of the individual. The 10k U.S. Adult Faces Database consists of over 10,000 images of faces of adults only that have a multitude of labels but include the relevant demographic labels needed for this project. These images are RGB oval-shaped close-ups of individual's faces with white-space surrounding the oval and are all 256 pixels tall and at most 256 pixels wide. For this dataset's race label, there are 6 categories, 5 are similar to the UTKFace dataset and the last extra category is Hispanics. These im-

ages were however pulled from Google Images and then labeled by survey takers. To provide a good label for the race, the mode race from the results of the surveys will be selected. Additionally, to make the images be all the same size for the convolutional neural network, these images width will be set to 256 pixels. Since the parts of these images that are not of a person's face are already whitespace, this transformation will be acceptable. Results for this project will be primarily be based on the gender classification error rate of each race for training with one convolutional neural network on all data and the error rate from the sequential two convolutional neural networks that separate by race and then predict gender. It is expected that the classification of gender will produce acceptable results in general but if the individual race performances are evaluated, the most common race in the datasets, White, will have the best accuracy. Ideally, the two convolutional neural network system will produce a result where the errors between races are closer than it was with one convolutional neural network. Evaluation of results will be based on the difference of errors. If the average error across all races can improve while the difference in errors between races goes down we will consider this to be good results. Certain cases in which the average error worsens by a small margin while the difference in error between races decreases will be considered and evaluated. There will be a secondary set of results to evaluated for this project as predictions of race need to be performed in order to get the final results. This will be evaluated with a simple examination of misclassification rates of the race for given images of faces. If the Gender Shades dataset becomes available, the results evaluation would change slightly as skin color, which is labeled in this dataset, will allow for a comparison of predicting gender based on similar skin color and race.

3. Technical Approach

In general, this project will be utilizing Python3 and PyTorch to load and transform data and to create the needed neural networks. To improve gender classification across all races, two convolutional neural networks will be used for the new proposed system. Both of these network's architectures will initially be based on the one that Levi & Hassner [3] proposed for age and gender classification using convolutional neural networks. Their network contains three convolutional layers, each followed by a rectified linear operation (ReLU) and a pooling layer, two fully-connected layers, each followed by a rectified linear operation (ReLU) and a dropout layer, and one fully-connected layer with an output fed into a soft-max layer. The decision to utilize this network was made because their proposed architecture was simple enough that it can be used when the amount of learning data is limited. This is key as the selected datasets at the moment do not have a considerable number of images and in the two convolutional neural networks pipeline, the images will be filtered by predicted race giving the second network fewer images to train on. Since the convolutional neural network proposed by Levi & Hassner was for age and gender classification, it will be examined and reworked for race classification. There are a minuscule number of publications utilizing convolutional neural networks to predict the race of individuals, therefore, their proposed architecture may need to be improved upon or different architecture may need to be developed. For the proposed pipeline, the intended way to pass the output from the first convolutional neural network to the second will be accomplished in a Pythonic way. After the first networks classify the race of an inputted test image, the predicted race label will then be utilized to pull a dataset that contains images that only have the same race label of what was predicted. If the dataset with labeled skin colors becomes available, this same method may be done except the predicted race and skin color labels will pull images that have either label matching rather than just race. With the filtered by race dataset specified, it will be used to train the second convolutional neural network to classify the gender of the earlier inputted test image.

4. Intermediate/Preliminary Results

For this milestone, an optimal convolutional neural network inspired by Levi & Hassner's proposal and the recommended pipeline still needs to be completed. For preliminary results, custom data loaders and image transformers were developed such that a convolutional neural network may predict the race or gender of an individual, but not both sequentially as proposed for my pipeline. The best gender and race classifying convolutional neural network in my implementation contained three convolutional layers, each followed by a ReLU and pooling layer, and two fully-connected layers with the last one feeding into a soft-max layer. This produced a max gender classification accuracy of 70.4%, which is considerably less than Levi & Hassner's best accuracy of 86.8% [3]. For race classification, performance was poor with a max race classification accuracy of 34.8%. I believe this low performance is due to the similarity of skin color for some races and due to the number of minors, especially infants in the UTKFace dataset. The similar color of races explanation for low accuracy is trivial but for minors and infants, I believe this is the case because facial features for people this age tend to be less pronounced than it would be when an individual is a young adult or older. Further examination on how the age has effects on prediction will need to be done. If low aged individuals continue to give problems in the classification of race, the dataset may have to be constricted to have no minors.

References

[1] W. Bainbridge, P. Isola, and A. Oliva. The intrinsic memorability of face photographs, 2013. In *Journal of Experimental Psychology: General*, pages 1323 - 1334.

[2] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification, 2018. In *Conference on Fairness, Accountability and Transparency*, pages 77–91.

[3] G. Levi and T. Hassner. Age and gender classification using convolutional neural networks, 2015. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) workshops*.

[4] Z. Zhang, Y. Song, and H. Qi. The intrinsic memorability of face photographs, 2017. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.