[THIS IS A CHECKPOINT FOR MY ONLINE REPO, NOT THE OFFICIAL PAPER] Can sampling based mitigation techniques lower gender bias within the MORPH-II dataset

1st Kaan Gögcay
Master of Applied Intelligence
Hogeschool van Amsterdam
Amsterdam, Netherlands

I. Introduction

In the Netherlands, the use of automated age prediction systems is becoming increasingly common. Over time, more supermarkets have begun integrating such systems into their checkout processes to automatically verify a customer's age for restricted purchases. Motivated by this growing trend, we, a group of Master's students in Applied Artificial Intelligence, aim to explore and improve the performance and fairness of age prediction models.

During our initial exploration, we examined several publicly available facial age datasets, including UTKFace, FG-NET, MORPH-II, and IMDB-WIKI. We observed that each dataset contains unique characteristics and varying levels of metadata quality. Among these, the MORPH-II dataset stood out due to its significant gender imbalance, with a substantial over-representation of male subjects. Such imbalances can lead to systematic differences in model performance across genders, resulting in gender bias.

Although this imbalance initially discouraged us from using MORPH-II for our main project, the dataset's quality and scale make it an excellent case for studying bias mitigation. Therefore, this research investigates whether sampling-based mitigation techniques can reduce gender bias in age prediction models trained on the MORPH-II dataset.

The central research question that guides this study is as follows: *To what extent can sampling-based mitigation techniques reduce gender bias in age prediction models trained on the MORPH-II dataset?*

II. Methods

The study consists of two main experimental workflows. One using the original dataset, and another using a dataset modified through sampling-based bias mitigation techniques. Each workflow follows the same procedure: exploratory data analysis (EDA), model training, evaluation, and measurement of gender bias. Ultimately, The models will be compared to determine whether the sampling mitigations lead to a reduction of gender bias.

A. Original Dataset

Exploratory data analysis of the MORPH-II dataset revealed several notable insights. The dataset contains a total of 50,015 images, making it larger than both UTKFace and FG-NET. Each image is accompanied by useful metadata, including age and gender. However, the dataset also has some limitations. The most prominent is the gender imbalance, with 86.1% male and 13.9% female subjects. Unlike UTKFace, MORPH-II does not provide information about ethnicity in its metadata. Additionally, MORPH-II contains multiple images per subject, which can lead to forms of data leakage or overfitting, as the model may encounter very similar faces across the training, validation, and test sets.

Experiments were conducted using different learning rates, numbers of epochs, and prediction tasks (classification versus regression). A learning rate of 0.001 was found to be too high, preventing effective learning. Reducing the learning rate to 0.0001 allowed the model to begin learning. However, overfitting occurred rapidly, after approximately six epochs, as illustrated in Figure 1.

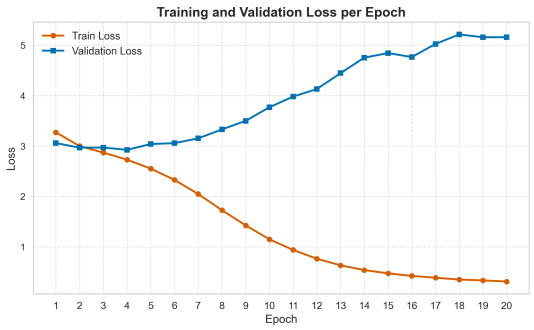

Fig. 1. Training and Validation Loss per Epoch.

My model has been trained using classification using 62 classes with the lowest class representing the age of 16 and the highest class 77. But as mentioned earlier, I also tested how well the model predicted using classification and using

regression. This workflow is copied from the DEX paper [1]. They also did this and got the lowest results compared to all other competing teams. After testing on the validation set with both the classification output and the regression output, it showed that regression always output a lower MAE score Figure 2. Therefore I will keep using this workflow of training with classification and predicting with regression. Also important to mention is that the model performed best after 4 epochs, since the MAE is lower than on other epochs Figure 2. Therefore I will use the model that is trained for 4 epochs for further testing

The model was trained using a classification approach with 62 classes, where the lowest class represents age 16 and the highest class represents age 77. Additionally, experiments were conducted to evaluate model performance using both classification and regression prediction strategies, following the workflow described in the DEX paper [1]. The DEX study reported that training using classification and predicing using regression resulted in the optimal performance, minimizing the final MAE score.

Evaluation on the validation set indicated that the regression outputs consistently achieved lower MAE scores than the classification outputs, as shown in Figure 2. Therefore, the workflow combining classification-based training with regression-based prediction was utilised. Additionally, the model trained for four epochs yielded the lowest MAE, outperforming models trained for other numbers of epochs, and was therefore selected for testing.

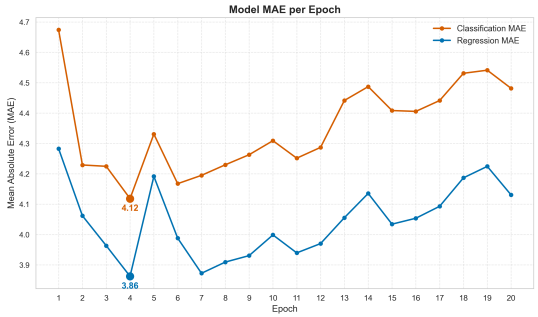

Fig. 2. MAE per epoch for classification and regression on validation set.

Finally, the mean absolute error (MAE) was evaluated separately for male and female subjects to assess whether the model differs in performance across genders. Given the significantly larger number of male samples in the dataset, it was hypothesized that the model would achieve lower MAE for male subjects compared to female subjects. This hypothesis is supported by the results shown in Figure 3, where the MAE for female subjects is slightly higher than for male subjects in both classification and regression predictions.

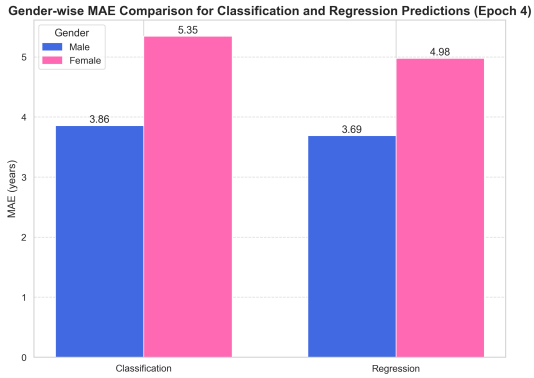

Fig. 3. MAE score comparison for men and women using classification and regression.

The magnitude of gender bias in the regression results can be quantified using the following formula:

$$\text{Bias}_{\text{abs}} = |\text{MAE}_{\text{male}} - \text{MAE}_{\text{female}}|$$

Applying this formula, the observed gender bias is 1.29. The objective in the next phase of the study is to reduce this bias through the application of sampling-based mitigation techniques.

B. *Sampling Mitigations*

- which mitigations i applied

C. *Modified/Mitigated Dataset*

- new eda - new modelling - new evaluations

D. *Comparison*

III. Results

IV. Discussions

- what are things i did wrong, questionable, improvement points, couldve done different

References

The following papers are relevant to this research and might be consulted throughout the study. These references are listed in the bibliography below. Note that my research won't be limited to these papers.

References

[1] R. Rothe, R. Timofte, and L. Van Gool, "Dex: Deep expectation of apparent age from a single image," in *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2015, pp. 252–257.