ISSN 1213-6972 (print)
ISSN 1213-6964 (online)
Journal of WSCG
http://www.wscg.eu
Vol.31, No-1-2, 2023

Bias mitigation techniques in image classification: fair machine learning in human heritage collections

Dalia Ortiz Pablo
Centre for Digital Humanities Uppsala
Department of ALM
Uppsala University
Sweden 75126, Uppsala
dalia.ortiz_pablo@abm.uu.se

Sushruth Badri **Erik Norén** **Christoph Nötzli**
Department of Information Techology
Uppsala University
Sweden 75126, Uppsala
{sushruth.badri.6580 | erik.noren.3194 |
christoph.notzli.7006}@student.uu.se

Abstract

A major problem with using automated classification systems is that if they are not engineered correctly and with fairness considerations, they could be detrimental to certain populations. Furthermore, while engineers have developed cutting-edge technologies for image classification, there is still a gap in the application of these models in human heritage collections, where data sets usually consist of low-quality pictures of people with diverse ethnicity, gender, and age. In this work, we evaluate three bias mitigation techniques using two state-of-the-art neural networks, Xception and EfficientNet, for gender classification. Moreover, we explore the use of transfer learning using a fair data set to overcome the training data scarcity. We evaluated the effectiveness of the bias mitigation pipeline on a cultural heritage collection of photographs from the 19th and 20th centuries, and we used the FairFace data set for the transfer learning experiments. After the evaluation, we found that transfer learning is a good technique that allows better performance when working with a small data set. Moreover, the fairest classifier was found to be accomplished using transfer learning, threshold change, re-weighting and image augmentation as bias mitigation methods.

Keywords

Image classification, Fairness, Bias mitigation, Gender classification, Transfer learning, Human Heritage Collection.

1 INTRODUCTION

Artificial intelligence (AI) systems have become a key instrument in many human decision processes. Their presence ranges from basic day-to-day tasks such as listening to music at home using technologies the size of a donut [McL19] to transcribing hand-written characters into machine-actionable text data [Cor20]. Those systems present clear benefits, mostly thanks to computers being able to perform tasks at a velocity that humans cannot and without getting tired. However, not all the outcomes are positive with AI. One major issue that those algorithms have presented is bias and unfairness. For example, the recruiting algorithm developed by Amazon, where the system learnt key traits from successful applicants' resumes to rate and find the top new candidate' CVs, exhibited a preference for males in technical positions [Kod19]. This problem does not only occur in contemporaneous sets but also historical scenarios.

Galleries, archives and museums carry deep insights into human memory and expression. Therefore not only collecting the remains of the past is relevant, but also analysing the objects that have been obtained. Furthermore, it is possible to apply and evaluate the technologies of the present, such as image classification, on the remnants of the past to understand how we can have more accurate and meaningful descriptions and classifications of heritage collections. Moreover, it is crucial to examine the ways in which qualitative aspects of human nature, such as bias, and the subjective nature of AI, intersect. This is especially relevant in the case of museum artifacts collections, which tend to exhibit inherent biases due to factors such as the methods of acquisition and digitization [Kiz21].

Literature on bias in AI mentions that bias is often encoded in the training data set and that this type of bias affects especially the models' accuracy [Par18]. However, even with well-balanced data sets, bias can be introduced in other steps of the ML pipeline. The work of

Wang et al. [Wan19] in gender classification is a good illustration of this. The authors showed that even when their data sets were perfectly balanced, the trained models resulted in biased predictions. Thus, assuming that the input data set is the only source and actor of bias is erroneous. In other studies, researchers have identified, named and classified many other sources of bias, for example, algorithmic bias, which appears due to inappropriate algorithmic choices, and is not present in the input data, or evaluation bias, which happens when the evaluation data set is not representative of the problem or the evaluation metric is incorrect for the task [Fah21, vanG22, Meh21].

The objective of this study is to investigate the effectiveness of algorithmic bias mitigation techniques for gender classification on a museum data set, thus bridging the gap between AI and gender classification in cultural heritage collections. Specifically, we aim to evaluate the performance of modern classification approaches and bias mitigation techniques on a highly diverse and gender-biased data set consisting of low-quality images of ancient people with varying ethnicity and ages. To the best of our knowledge, no studies have yet implemented these techniques for this purpose.

2 RELATED WORK

2.1 Bias mitigation

Technical bias mitigation techniques can be divided into five stages: problem understanding, preprocessing, in-processing, post-processing and deployment [Hor22, Bel18]. Bias mitigation methods performed on the training data are considered preprocessing; techniques performed while training ML models are categorised as in-processing; and post-processing methods are applied to trained ML models [Hor22, Bel18].

One of the most common techniques for bias mitigation in the pre-processing step is data augmentation - a technique used to increase the amount of data in the training set by performing modifications (e.g. rotations, colour transformation, reflection) on the original set [Mij18]. In the work done by McLaughlin et al. [McL15], the authors evaluated the effectiveness of different data augmentation techniques in re-identification systems. They found that changing an image background increases the performance of a model only when a combination of other augmentation techniques, such as cropping and mirroring, are also used.

Other bias mitigation methods have also been investigated in the literature. The work of Wang et al. [Wan20] evaluates strategic re-sampling, adversarial training, domain discriminative training, and domain-independent training in a gender classification scenario using CNNs and a data set composed of pictures of celebrities. The authors found that oversampling outperformed the other techniques, and that was followed closely by domain-independent training. Similarly, Lee et al. [Lee22] revise bias mitigation methods for CNNs, e.g. ReBias and vanilla, to create a benchmark for the bias mitigation pipeline. The study showed that state-of-the-art approaches achieved different approaches depending on the training data set, which suggests that a bias mitigation process is task specific.

2.2 The FairFace data set

Neural network models have been shown to learn and amplify biases in training data [Hal22]. This is partly the motivation for the creation of the FairFace data set, presented in the work by Kärkkäinen and Joo [Kar19]. The FairFace data set contains 108 501 images, and it is balanced concerning gender, ethnicity, and age, and therefore does not suffer from the same bias that other big data sets do. In [Kar19], it is shown that models trained with the FairFace data set generalize better than other existing face data sets to unseen and ethnically diverse data. In the work of Kotti et al. [Kot22], the data set is used in their experiments for evaluating bias in Generative Adversarial Networks (GANs). Moreover, a model trained on the FairFace data set was used in [Dev22] in order to benchmark the performance of their new fair model. In our work, we conduct transfer learning using the FairFace data set for the models to learn first for a known fair data set.

3 BACKGROUND

3.1 Deep learning models

Our classifiers are based on existing networks provided in the tensorflow framework. We chose Xception [Chol17] and EfficientNet [Tan19] as base networks. Both were implemented using Keras' model API.

The Xception network was first introduced in 2017 in the paper "Xception: Deep Learning with Depthwise Separable Convolutions" by Chollet [Chol17]. The model is a fully convolutional network designed with the goal of being a more efficient variant of the Inception architecture from 2014 [Sze15]. The technique that separates the Xception architecture from the Inception architecture is the implementation of depthwise separable convolutions instead of regular convolutions. Regular convolutions work by performing convolutions on all channels at once, unlike depthwise separable convolutions which performs a single convolution operation on each input channel. The Xception architecture has 36 convolutional layers and has an input size of 299x299.

EfficientNet was introduced in the 2019 paper "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks" [Tan19]. Rather than being a single network, EfficientNet is best described as a family of

networks architectures, created from scaling the baseline network EfficientNet-B0. The paper introduces a new approach to scaling models which included scaling the network's width, depth and image resolution together. The EfficientNet architecture that we implemented was EfficientNet-B3 which has an input size of 300x300 which is similar to Xception. In [Tan19], the EfficientNets' performances, compared to other convolutional neural networks, is the state-of-the-art of several data sets, while reducing the size of the models. The EfficientNet-B3 architecture is about half the size of Xception, with 12 million versus 23 million parameters.

Both the models have the same four top layers appended to produce the binary classification. These consists of a GlobalAveragePooling2D layer, a BatchNormalization layer, a Dropout layer with a dropout rate of 0.2, and lastly, a dense layer for binary classification with a sigmoid activation function [Keras].

3.2 Fairness metrics

Bias quantification metrics are closely linked to the concept of *fairness*, which has two main definitions. The first definition is individual fairness, which involves assessing similar individuals and expecting them to be treated similarly by the model [Cho17]. The second definition is group fairness, which refers to the absence of prejudice and favoritism towards a particular group [Meh21]. Further, there are two common categories of fairness metrics related to the previous definitions: *Definitions Based on Predicted Outcome* and *Definitions Based on Predicted and Actual Outcome*. Some examples of fairness metrics that belong to these categories are:

- Demographic (or statistical) Parity Difference (DPD), which focuses on ensuring that there are similar amounts of positive predictions across groups. In this context, a classifier is called fair if [Ver18]:
$$\mathrm{TP}+\mathrm{FP}=\mathrm{TN}+\mathrm{FN} \tag{1}$$
- Proportional Parity Difference (PPD) is a normalized version of DPD [Koz21]. In the binary case a classifier is called fair if:
$$\frac{\mathrm{TP}+\mathrm{FP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}}=\frac{\mathrm{TN}+\mathrm{FN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}} \tag{2}$$
In a problem with more than two classes, the denominators in equation 2 would be different, resulting in different scaling factors for each class. However, in the binary case, the denominators are the same. In our work, we treated them as different because we did not explicitly account for the normalization constant in any of the experiments.
- Equality of Opportunity. In a binary classification task, it aims to ensure that both groups have equal rates of true positives. Definitions are based on Predicted and Actual Outcome, and in this case a classifier is fair if [Gar20]:
$$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}=\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \tag{3}$$
- Predictive Rate Parity Difference. This measure, in a binary classification task, ensures that both groups have equal rates of predicted positives [Cho17]. The definition is based on Predicted and Actual Outcome [Ver18]. A classifier is fair if: In the binary case:
$$\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}=\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FN}} \tag{4}$$

This metrics are popular in the general area of fair machine learning and can be applied in image classification as well. Examples of this are Demographic Parity used in [Fra21] and Equality of Opportunity used in [Sto22, Zha18]. Therefore, in this project, we use demographic parity difference, equality of opportunity, proportional parity difference and predictive rate parity difference as fairness metrics.

3.3 Bias mitigation techniques

3.3.1 Reweighting

Reweighting is a pre-processing bias mitigation method. The idea of this method is to give classes that are more common in the training data set a lower weight i.e. the sample has less effect on the training of the model. Reweighing approach also maintains a high accuracy level [Kam12]. Pre-processing techniques try to transform the data so that the underlying discrimination is removed [Meh21]. The class weights are assigned and passed to the model while training through Keras implementation.

3.3.2 Image augmentation

Image augmentation is an in-processing method for bias mitigation. In this project, data augmentation is implemented by altering the training data in order to deal with classification bias in under-representation of certain groups. Data augmentation can reduce classification error for discriminated groups. Furthermore, even though different classifiers do not perform equally good, they exhibit positive results when data augmentation takes place [Ios18]. In our project the data augmentation has four pre-processing layers that are added at the beginning of the model. These layers perform random flip, random rotation, random translation (i.e. random movement), and random contrast. The pre-processing layers are implemented using the Keras API for image augmentation layers [Keras].

3.3.3 Threshold change

Changing the threshold of the model is a post-processing method [Hor22]. In the standard implementation of our model we use a threshold of 0.5 i.e. every predicted value below 0.5 is interpreted as Female. This threshold is not necessarily optimized for fairness. We applied the following threshold changes to the models:

Equal true In this case, we optimize the threshold to the minimal difference of predicted true values in each class:

$$\min\{|\mathrm{TP}_t - \mathrm{TN}_t|\} \quad \text{with t} = [0,1], \tag{5}$$

where t, here and in the following definitions, is the value of the threshold for the classifier.

Equal false In this case we optimize the threshold to the minimal difference of predicted false values in each class.

$$\min\{|\mathrm{FP}_t - \mathrm{FN}_t|\} \quad \text{with t} = [0,1] \tag{6}$$

Equal total In this case we optimize the threshold to predict the minimal difference of predicted values in each class.

$$\min\{|(\mathrm{TP}_t + \mathrm{FP}_t) - (\mathrm{TN}_t + \mathrm{FN}_t)|\} \quad \text{with t} = [0,1] \tag{7}$$

Equal opportunity In this case we optimize the threshold to predict the minimal difference of predicted values in each class.

$$\min\{|\frac{\mathrm{TP}_t}{\mathrm{TP}_t + \mathrm{FN}_t} - \frac{\mathrm{TN}_t}{\mathrm{TN}_t + \mathrm{FP}_t}|\} \quad \text{with t} = [0,1] \tag{8}$$

3.3.4 Transfer Learning

Transfer learning is the ML technique of taking the knowledge that is able to be learned by training a model on one task and then fine tuning it to a different but related task [Pan10]. In this project, transfer learning was implemented as a way overcome training data scarcity. For this part of the process, the FairFace data set is the source domain and the cultural heritage data set is the target domain. The tasks are similar in both domains, being binary image classification of gender in both cases.

Through our experiments, it will be investigated if it is possible to transfer a fair model trained on the FairFace data set. This will be done by comparing the results of the models implementing transfer learning against the ones trained on only target domain. The metrics for evaluating bias will be fairness metrics detailed in section 3.2 and section 3.3.

3.4 Cultural Heritage Data Set

The Cultural Heritage Data Set (CHDS) contains labelled images from the The National Museums of World Culture [VKM] in Sweden, and the data set was manually post-processed as part of an earlier part of the Quantifying Culture project. Within the data set there are 3128 images in total of people, where 1067 images are labeled male, and 2061 labeled female (a sample of the data set is shown in Figure 1). These images are photographs taken "in-the-wild" and do not guarantee that they contain only one person at a time, but ensure that all the individuals are of the same sex. The time period which the photographs stem from are either the nineteenth or twentieth century. After performing the face extraction described in section 4.1, 4897 faces were detected, 2331 labeled male, and 2566 labeled female.

Figure 1: Example of original image from the Cultural Heritage Data Set [VKM] containing two individuals labeled males.

4 IMPLEMENTATION DETAILS

4.1 Face extraction

Since the CHDS contains full-body images of people, and our model is designed to perform gender classification based on facial features, face extraction is performed on the CHDS. The face extraction module makes use of the Python API of the dlib ML toolkit [Kin09]. Within the module a pre-trained CNN model is used for face detection which then allows for extracting cropped images of the faces contained in the photographs of the CHDS.

4.2 Experimental setup

We use three experimental setups to test our networks. In the first experiment, we used only the CHDS to train and test our networks. In a second step, we used the FairFace data to pretrain validate and test them our models. The third experiments use the FairFace models from the second experiments for transfer learning. The transfer learned models are then validated, tested

and compared to the baseline models. For all experiments the fairness metrics discussed in section 3.2 are implemented for evaluation.

4.2.1 *Baseline CHDS experiments*

In the baseline model experiments we train and cross validate the models shown in section 3.1. For each of these two models, there are three variations tested, which depend on the bias mitigation method being evaluated. The first variation is without any bias mitigation method; the second is with class re-weighting applied; and the third is with data set augmentation applied. These bias mitigation methods are covered in section 3.3. The models were trained for a maximum of 20 epochs with the learning rate decreasing exponentially. We added an early stopping to the training, which means that the training will stop in case the training does not decrease the validation loss in three consecutive epochs.

4.2.2 *FairFace experiments*

The models described in section 3.1 are pretrained with the help of the FairFace data in three different ways. First we just train the models, in a second experiment we reweight the classes because there is a slight bias towards the Male class in the FairFace data. In the third experiment we add the augmentation layers described in section 3.3. These experiments are run to create the different variations of base models to be used for transfer learning, as well as a way to evaluate performance on the models on a data set that we know to be fair and balanced.

4.2.3 *Transfer learning experiments*

To build some the classifiers for this project, we train a base model, either EfficientNet-B3 or Xception, with the FairFace training data. During training, the weights of the models' layers are tuned and adjusted in order to increase accuracy. This trained model is then moved to the target domain and has a number of its layers and their parameters frozen in order to keep the knowledge learned in the source domain. For transfer learning with the EfficientNet-B3 model all layers except the the last forty layers are frozen. In the Xception case all layers except the last ten layers and the four top layers are frozen. Also, we compare four different version of the transfer learned models. We use the unweighted and unaugmented pretrained model to test transfer learning.

4.2.4 *Data set splits*

As mentioned in section 3.4 the cultural heritage data set does not contain a lot of data. Therefore we choose to use 80% of the data for training, 10% for validation and 10% for testing. We used the same splits for the training with the help of the FairFace data set. For cross validation we used a 5-Fold cross validation split, 80% of data for training and 20% for validation.

4.2.5 *Early stopping*

We use early stopping provided by Keras [Keras]. I.e. the training stops when the validation loss is not reduced in three consecutive training steps.

4.2.6 *Used infrastructure for training*

For the training of the models we used the Alvis cluster of Chalmers University [Alvis]. We used the A100 GPUs of the cluster. This allowed us to use larger batch sizes and smaller training time.

5 RESULTS

The results of our experiments can be found in this section. Within table 1 the results of the Baseline CHDS experiments and the Transfer learning experiments can be seen. Positive values in the fairness metrics Demographic Parity Difference, Proportional Parity Difference and Predictive Rate Parity Difference show a bias in favor of the Male class. In the Equality of Opportunity metric the bias is in favor of the Male class if the values are negative.

5.1 Baseline CHDS experiments

Baseline CHDS experiments were performed for both the EfficientNet-B3 and Xception models. Figure 2 shows the performance in accuracy for EfficientNet and Xception. It can be seen that all three variations of the EfficientNet model reaches higher accuracy than the Xception variations. Figure 3 shows the training and validation accuracies for the Xception variations. The EfficientNet variations also have greater performance in the fairness metrics as well as seen in table 1.

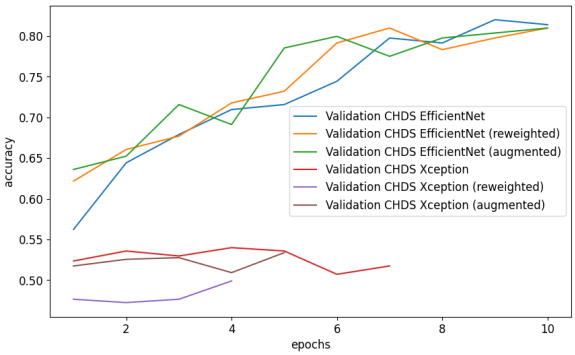

Figure 2: EfficientNet and Xception validation accuracy for the CHDS experiments.

5.2 FairFace experiments

The model that achieves the highest performance in accuracy on the FairFace data set is EfficientNet with a validation accuracy of ∼91% as seen in Figure 4. The best performing Xception model is the one implemented with augmentation which achieves a validation accuracy of ∼87%. Figure 5 shows the performance of the models with regards to Demographic Parity Difference where all variations except Xception with

Network name	Transfer learning	Threshold change	Reweighting	Image augmentation	Accuracy	Demographic Parity Difference	Proportional Parity Difference	Equality of Opportunity	Predictive Rate Parity Difference
EfficientNet	No	No	No	No	78.5 +/- 1.2%	-39 +/- 54.6	-0.04 +/- 0.06	0.01 +/- 0.06	-0.03 +/- 0.05
EfficientNet	No	Equal false	No	No	78.9%	4	0.008	0.036	-0.053
EfficientNet	No	No	Yes	No	80.4 +/- 1.6%	-64 +/- 33.8	-0.06 +/- 0.03	0.04 +/- 0.03	-0.01 +/- 0.03
EfficientNet	No	Equal total	Yes	No	79.9%	6	0.012	0.041	-0.056
EfficientNet	No	No	No	Yes	78 +/- 2.2%	-155 +/- 77.8	-0.16 +/- 0.08	0.13 +/- 0.09	0.04 +/- 0.06
EfficientNet	No	Equal opp.	No	Yes	80.1%	-8	-0.016	0.013	-0.039
EfficientNet	Yes	No	No	No	84.5 +/- 0.5%	-45 +/- 42.3	0.05 +/- 0.04	0.01 +/- 0.04	0.02 +/- 0.03
EfficientNet	Yes	Equal false	No	No	83.7%	-8	-0.016	0.017	-0.038
EfficientNet	Yes	No	No	Yes	83.6 +/- 0.3%	-115 +/- 26.2	-0.11 +/- 0.02	-0.08 +/- 0.04	0.03 +/- 0.04
EfficientNet	Yes	Equal false	No	Yes	82.1%	12	0.024	0.056	-0.064
EfficientNet	Yes	No	Yes	No	84.2 +/- 1.5%	-101 +/- 73.5	-0.1 +/- 0.07	0.07 +/- 0.08	0.02 +/- 0.06
EfficientNet	Yes	Equal false	Yes	No	80.7%	14	0.028	0.059	-0.066
EfficientNet	Yes	No	Yes	Yes	83.9 +/- 1.4%	-13.4 +/- 48.2	-0.01 +/- 0.05	0.02 +/- 0.04	-0.04 +/- 0.03
EfficientNet	Yes	Equal total	Yes	Yes	82.1%	0	0	0.031	-0.049
Xception	No	No	No	No	52.4%	-492	-1	-1	-0.524
Xception	No	No	Yes	No	47.6%	492	1	1	0.476
Xception	No	No	No	Yes	52.4%	-492	-1	-1	-0.524
Xception	Yes	No	No	No	65.6 +/- 3.7%	-7 +/- 435	-0.01 +/- 0.44	-0.01 +/- 0.45	-0.06 +/- 0.15
Xception	Yes	Equal total	No	No	78.3%	2	0.004	0.032	-0.051
Xception	Yes	No	No	Yes	72.4 +/- 0.1%	-338 +/- 401	-0.34 +/- 0.40	-0.33 +/- 0.41	0.1 +/- 0.16
Xception	Yes	Equal opp.	No	Yes	80.9%	-16	-0.033	-0.002	-0.029
Xception	Yes	No	Yes	No	66.4 +/- 6%	-341 +/- 398	-0.35 +/- 0.41	-0.34 +/- 0.41	0.09 +/- 0.19
Xception	Yes	Equal opp.	Yes	No	77.4%	-14	-0.028	-0.002	-0.033
Xception	Yes	No	Yes	Yes	73.1 +/- 7.8%	-389 +/- 169	-0.39 +/- 0.17	-0.37 +/- 0.19	0.15 +/- 0.08
Xception	Yes	Equal opp.	Yes	Yes	81.3%	4	0.008	0.039	-0.054

Table 1: Validation results of the experiments. The results of the CHDS experiments are the entires marked with "No" in the transfer learning column. Entries marked "Yes" in the Transfer learning column are implemented with transfer learning from the FairFace data set. In the table the performance of different combinations of the different bias mitigation methods can be seen.

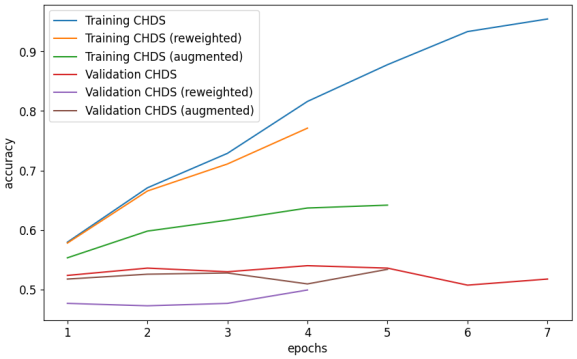

Figure 3: Xception training and validation accuracy for the CHDS experiments.

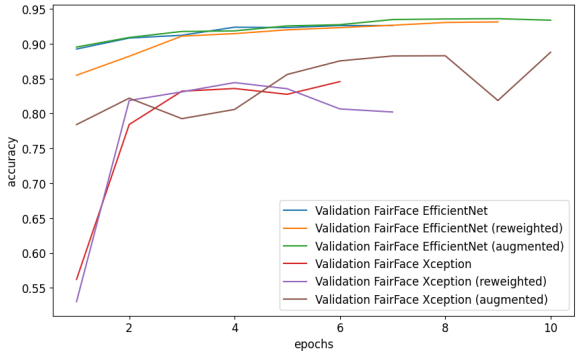

Figure 4: EfficientNet and Xception validation accuracy for the FairFace experiments.

reweighting have similar results. This is also the case when evaluating with regards to Equality of Opportunity Difference, as seen in Figure 6.

5.3 Transfer learning experiments

Using the models created in the FairFace experiments for transfer learning, the accuracy results shown in table 1 were achieved. EfficientNet achieves the highest accuracy of the transfer learned models with 83.7%.

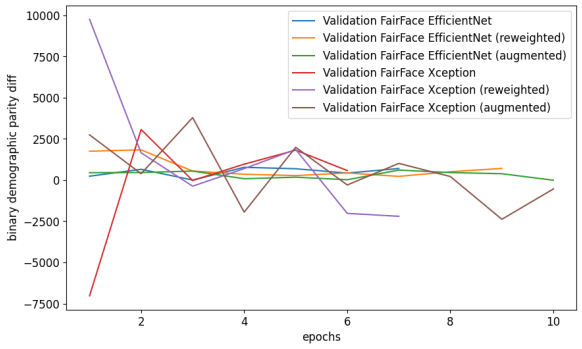

Figure 5: EfficientNet and Xception validation Demographic Parity Difference performance for the FairFace experiments.

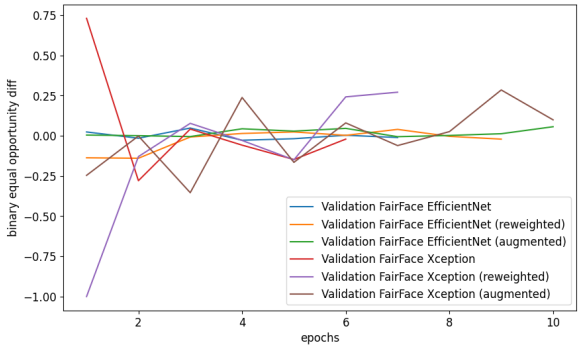

Figure 6: EfficientNet and Xception validation Equality of Opportunity performance for the FairFace experiments.

In the fairness metrics different transfer learning variation of both EfficientNet and Xception have the best performance. Measured in Demographic Parity Difference the EfficientNet model with equal total threshold change, re-weighting, and image augmentation performs best. If Equality of Opportunity is considered instead the re-weighted Xception model is best performer. When using augmentation and the EfficientNet network the bias metrics stay close to 0 during the training as seen in Figure 8 and Figure 7.

6 DISCUSSION

The Xception network is not ideal for a use case with a small data set, in this case the CHDS. In table 1 the results show that the Xception network is not performing well due to overfitting as seen in Figure 3. More regularization within the network and the final layers could be a possible solution to fix the overfitting in the base case of the Xception model. However, here the value of transfer learning is shown to be an effective way to improve a model which poor performance is partly due to lack of training data. Xception used with transfer learning allowed for a validation accuracy of 73.1%, which greatly outperforms the baseline CHDS trained models as seen in Table table 1.

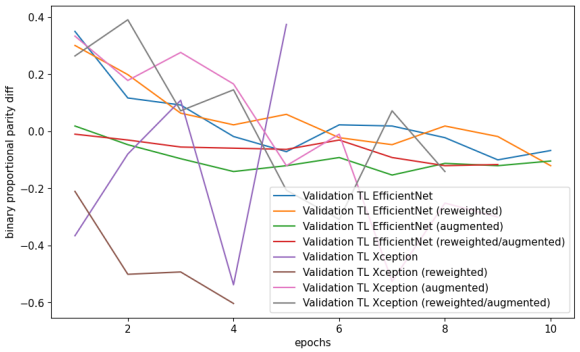

Figure 7: EfficientNet and Xception validation Proportional Parity Difference for the transfer learning experiments.

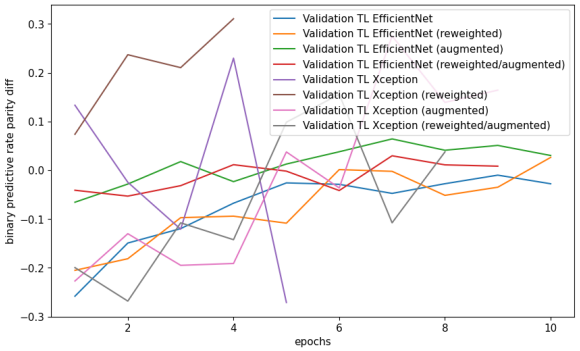

Figure 8: Xception training and validation Predictive Rate Parity Difference for the transfer learning experiments.

The results in section 5 show that EfficientNet has a higher accuracy and is better than Xception in regard to fairness in the transfer learning variations. These results are expected since EfficientNet is a newer network and also shows better performance in the benchmarks [Tan19].

Using transfer learning to train the models is a good choice in our use case. The biggest advantage is that we get a higher accuracy due to training the models with more data. Comparing the EfficientNet model, with or without transfer learning, it can be seen in table 1 that accuracy in increased from 78.5% to 84.5%. However, as a method for bias mitigation transfer learning does not make much of an impact. For the experiments on the EfficientNet variations the fairness metrics do not appear to be influenced by applying transfer learning or not. Due to the poor performance of the baseline variations of Xception, it is not possible to see if previous results translate to the Xception model.

When transfer learning and augmentation are used together with reweighting on EfficientNet, the fairness metrics are improved. This is also the case for accuracy as seen in table 1. Using reweighting alone does not show any significant improvement in the metrics for both EfficientNet and Xception. It could be due to the

fact that there is not a big difference in the amount of samples in each of the classes in CHDS.

Comparing the EfficientNet CHDS results in table 1, with and without image augmentation, it can be seen that accuracy is not changed much, the accuracy score being 78.5% +/- 1.2% without and 78% +/- 2.2% with. Moreover, when evaluating these same two variations by the fairness metrics no apparent improvement can be seen in any case. This pattern emerges in the transfer learning experiments as well for EfficientNet. Using image augmentation alone slightly worsens performance in terms of accuracy and improvement cannot be seen in the fairness metrics either. However, when comparing the transfer learning variations, with and without image augmentation, with reweighting implemented, the pattern does not repeat. In this case image augmentation improved performance in all accord with all fairness metrics while reaching a perfect score for Demographic- and Proportional Parity. For transfer learning experiments for Xception, image augmentation improves the fairness metrics when paired together with reweighting, similarly to EfficientNet. Image augmentation improves accuracy for the Xception variations. One important observation is that augmentation stabilizes the training (metrics always close to the ideal value) for EfficientNet as seen in Figure 8 and Figure 7. This is advantageous because we can stop the training at any point with similar bias metric values.

When applied, threshold change improves performance in the fairness metrics consistently. In all EfficentNet experiments the improvement in fairness came at the cost of accuracy, this is however not translated in the Xception experiments. It could be because of Xception network overfitting on the data and it favours a certain class during prediction. This can be seen from the metric values in 1. When using threshold change, some of the wrong predictions move over to the other side of the decision boundary and so accuracy improves.

Regarding the CHDS data set, it shows a slight bias favouring the Female class (52.4% of the images in the Female class). This is also the case for our overall best model (EfficientNet, with reweighting and image augmentation but no thresholding). For example, the value of bias according to the Proportional Parity Difference is 1 +/- 5% towards the Female class. Gender is something that should be considered concerning fairness; however, there is a cause to consider fairness in relation to other metrics as well. With the CHDS data set only having gender as its label, further investigation into other sources of bias is difficult to attain. The CHDS data set is diverse concerning age, ethnicity, culture, and the period the photo was taken. All of these aspects add the possibility for bias. Photographs portraying a given group of people of a certain ethnicity could have been taken in poorer conditions than photographs of another group due to the time period or other external conditions. As a result, this could lead to it being harder for the model to learn to classify the first group correctly, which might not show up in our results evaluated by gender.

Due to using a pre-trained model in the face extraction module for the cultural heritage data set (section 3.4), there is an additional potential source of bias. For example, it is possible that the pre-trained model was itself trained with biased data, thus leading to it potentially having a better ability to extract the faces of a certain group. Since we did not investigate the total amount of faces belonging to each class in the photographs in the CHDS, no evaluation of the potential bias of the face extraction module was performed.

7 CONCLUSION

In this paper, we evaluated three novel bias mitigation techniques for image classification in a cultural heritage data set. We found that individually implementing augmentation, class re-weighting, and threshold change does not lead to a fairer model compared to the baseline model. We have also evaluated the performance of the classifiers when implementing transfer learning. We have shown that, for this task, combining transfer learning with image augmentation, class re-weighting, and threshold change is the best way to reach a fair classifier.

8 FUTURE WORK

In our current implementation the process of fine tuning was not included in the transfer learning implementation. In fine tuning the final models are trained, with all layers unfrozen, with a very low learning rate for just a few epochs. This could potentially improve performance, mainly accuracy.

Further work could be done in tuning the hyperparameters of our models. Since the priority of this project was the evaluation of the bias mitigation methods, less focus was put on perfecting model performance. Therefore work can be done in investigating optimal learning rate, batch size, weight decay etcetera.

As discussed in section 6, improvement to our analysis could be done if we would have had access to multiple labels, e.g. age, ethnicity, etcetera. With more labels we could take a look a the bias from an intersectional stand point. With the help of intersectionality we could show further biases in the approach, and we could take action to reduce these biases.

Another interesting addition to the project would be to compare the results of the EfficentNet with an implementation of the EfficientNetV2. EfficientNetV2 includes data augmentation layers and we had promising preliminary results for the base model.

9 ACKNOWLEDGMENTS

This work has been partially supported by the WASP-HS (Autonomous Systems and Software Program-Humanities and Society) grant - an initiative of the Wallenberg Foundations; project title: "Quantifying Culture: AI and Heritage Collections"; project number: MAW 2020.0054. The computations/data handling were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at Chalmers Centre for Computational Science and Engineering (C3SE) partially funded by the Swedish Research Council through grant agreement no. 2018-05973, project Dnr: SNIC 2022/22-1091.

10 REFERENCES

[All22] Allawala, A., Ramteke, A., & Wadhwa, P. Performance Impact of Minority Class Reweighting on XGBoost-based Anomaly Detection. International Journal of Machine Learning and Computing, 12(4), 2022.

[Alvis] "Chalmers Centre for Computational Science and Engineering": https://www.c3se.chalmers.se/about/Alvis/.

[Bel18] Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Zhang, Y. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:1810.01943, 2018.

[Cho17] Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1251-1258), 2017.

[Cho17] Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data, 5(2), 153-163, 2017.

[Cho20] Chouldechova, A., & Roth, A. A snapshot of the frontiers of fairness in machine learning. Communications of the ACM, 63(5), 82-89, 2020.

[Con00a] Conger., S., and Loch, K.D. (eds.). Ethics and computer use. Com.of ACM 38, No.12, 2000.

[Cor20] Cordell, R. Machine learning and libraries: a report on the state of the field. Library of Congress, 2020.

[Dev22] Deviyani, A. Assessing Dataset Bias in Computer Vision. arXiv preprint arXiv:2205.01811, 2022.

[Dro20] Drozdowski, P., Rathgeb, C., Dantcheva, A., Damer, N., & Busch, C. Demographic bias in biometrics: A survey on an emerging challenge. IEEE Transactions on Technology and Society, 1(2), 89-103, 2020.

[Fah21] Fahse, T., Huber, V., & van Giffen, B. Managing bias in machine learning projects. In Innovation Through Information Systems: Volume II: A Collection of Latest Research on Technology Issues (pp. 94-109). Springer International Publishing, 2021.

[Fra21] Franco, D., Oneto, L., Navarin, N., & Anguita, D. Learn and Visually Explain Deep Fair Models: an Application to Face Recognition. In 2021 International Joint Conference on Neural Networks (IJCNN) (pp. 1-10). IEEE, 2021.

[Gar20] Garg, P., Villasenor, J., & Foggo, V. Fairness metrics: A comparative analysis. In 2020 IEEE International Conference on Big Data (Big Data) (pp. 3662-3666). IEEE, 2020.

[Hal22] Hall, M., van der Maaten, L., Gustafson, L., & Adcock, A. A systematic study of bias amplification. arXiv preprint arXiv:2201.11706, 2022.

[Hor22] Hort, M., Chen, Z., Zhang, J. M., Sarro, F., & Harman, M. Bias mitigation for machine learning classifiers: A comprehensive survey. arXiv preprint arXiv:2207.07068, 2022.

[Ios18] Iosifidis, V., & Ntoutsi, E. Dealing with bias via data augmentation in supervised learning scenarios. Jo Bates Paul D. Clough Robert Jäschke, 24, 11, 2018.

[Jin20] Jin, X., Barbieri, F., Davani, A. M., Kennedy, B., Neves, L., & Ren, X. Efficiently mitigating classification bias via transfer learning. arXiv preprint arXiv:2010.12864, 2020.

[Kam12] Kamiran, F., & Calders, T. Data preprocessing techniques for classification without discrimination. Knowledge and information systems, 33(1), 1-33, 2012.

[Kar19] Kärkkäinen, K., & Joo, J. Fairface: Face attribute dataset for balanced race, gender, and age. arXiv preprint arXiv:1908.04913, 2019.

[Keras] Keras, C. F. Theano-based deep learning libraryCode: https://github.com/fchollet. Documentation: http://keras.io, 2015.

[Kin09] King, D. E. Dlib-ml: A machine learning toolkit. The Journal of Machine Learning Research, 10, 1755-1758, 2009.

[Kiz21] Kizhner, I., Terras, M., Rumyantsev, M., Khokhlova, V., Demeshkova, E., Rudov, I., & Afanasieva, J. Digital cultural colonialism: measuring bias in aggregated digitized content held in Google Arts and Culture. Digital Scholarship in the Humanities, 36(3), 607-640, 2021.

[Kod19] Kodiyan, A. A. An overview of ethical issues in using AI systems in hiring with a case study of Amazon' s AI based hiring tool. Researchgate Preprint, 1-19, 2019.

[Kot22] Kotti, S., Vatsa, M., & Singh, R. On Biased Behavior of GANs for Face Verification. arXiv

preprint arXiv:2208.13061, 2022.

[Koz21] Kozodoi, N., & Varga, T. V. fairness: Algorithmic Fairness Metrics. URL https://CRAN. R-project. org/package= fairness. R package version, 1(1), 228, 2021.

[Lee22] Lee, J., Lee, J., Jung, S., & Choo, J. DebiasBench: Benchmark for Fair Comparison of Debiasing in Image Classification. arXiv preprint arXiv:2206.03680, 2022.

[McL15] McLaughlin, N., Del Rincon, J. M., & Miller, P. Data-augmentation for reducing dataset bias in person re-identification. In 2015 12th IEEE International conference on advanced video and signal based surveillance (AVSS) (pp. 1-6). IEEE, 2015.

[McL19] McLean, G., & Osei-Frimpong, K. Hey Alexa... examine the variables influencing the use of artificial intelligent in-home voice assistants. Computers in Human Behavior, 99, 28-37, 2019.

[Meh21] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR), 54(6), 1-35, 2021.

[Mij18] Mikołajczyk, A., & Grochowski, M. Data augmentation for improving deep learning in image classification problem. In 2018 international interdisciplinary PhD workshop (IIPhDW) (pp. 117-122). IEEE, 2018.

[Pan10] Pan, S. J., & Yang, Q. A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10), 1345-1359, 2010.

[Par18] Park, J. H., Shin, J., & Fung, P. Reducing gender bias in abusive language detection. arXiv preprint arXiv:1808.07231, 2018.

[Sto22] Stone, R. S., Ravikumar, N., Bulpitt, A. J., & Hogg, D. C. Epistemic Uncertainty-Weighted Loss for Visual Bias Mitigation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2898-2905), 2022.

[Sze15] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9), 2015.

[Tan19] Tan, M., & Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In International conference on machine learning (pp. 6105-6114). PMLR, 2019.

[vanG22] van Giffen, B., Herhausen, D., & Fahse, T. Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. Journal of Business Research, 144, 93-106, 2022.

[Ver18] Verma, S., & Rubin, J. Fairness definitions explained. In Proceedings of the international workshop on software fairness (pp. 1-7), 2018.

[VKM] "The National Museums of World Culture": https://www.varldskulturmuseerna.se/en/.

[Wan19] Wang, T., Zhao, J., Yatskar, M., Chang, K. W., & Ordonez, V. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 5310-5319), 2019.

[Wan20] Wang, Z., Qinami, K., Karakozis, I. C., Genova, K., Nair, P., Hata, K., & Russakovsky, O. Towards fairness in visual recognition: Effective strategies for bias mitigation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8919-8928), 2020.

[Zha18] Zhang, B. H., Lemoine, B., & Mitchell, M. Mitigating unwanted biases with adversarial learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 335-340), 2018.