ALGORITHMIC FAIRNESS - FAIRNESS IN COMPUTER VISION

Victor Célérier

s201836@student.dtu.dk
Technical University of Denmark
Lyngby

Zineb Fadili

s201501@student.dtu.dk
Technical University of Denmark
Lyngby

1. INTRODUCTION

As artificial intelligence is more and more important in everyday choices, fairness has become a major concern for many scientists and researchers. It is evident by the increase of papers related to this subject that have been published in the past three years.
In machine learning, a model is said to be fair when its output is not influenced by protected attributes such as race, gender... As the decisions taken by artificial intelligence spread from granting a loan to hiring employees, a bias in these choices can tremendously affect society. There have been multiple scandals in the past where big companies such as Amazon, Facebook, and Apple had machine learning models exhibiting biases towards protected attributes. For instance, Amazon discovered in 2015 that its hiring system was not gender-neutral.
However, if alleviating bias in machine learning is a priority, keeping a satisfying accuracy and a low error is also important. Also, diverse metrics exist to assess the fairness of a machine learning model and these metrics don't evaluate the same criteria and have different degrees of strength.
In this project, we built a classifier determining the attractiveness of people based on pictures of their faces. Through this study, we wanted to understand what each of these metrics is measuring, what their limits are, and how we can use them to alleviate the partiality manifested by our neural network. This study is also an opportunity to test different solutions to reduce our model's bias.
After building and training the classifier, we tested the effect of a postprocessing tool, threshold adjustment, and a preprocessing one, dataset balancing.

2. SOURCE LINKS:

- Github repository with source code and poster:
  https://github.com/zinebfadili/computer-vision-fairness
- Link to the trained models:
  https://drive.google.com/drive/folders/1dqvR1g0aRgrdphttaGt2W-4UTqcxeO9k?usp=sharing
- Dataset:
  http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

3. RELATED WORK

In the past few years, many researchers developed an interest in fairness in machine learning. As racial and gender injustice discussions have become more present in popular debates, this has transpired into the topic of algorithmic fairness. Concerning image classification for the celebA dataset, the one we will be using in this project, multiple small independent and bigger research projects have worked on it. The paper [1] compares diverse classifier used on this dataset. We will be using this article in a later section to evaluate the performance of our neural network.
Although the classifiers in this paper don't aim specifically at attractiveness, they were still good benchmarks to assess our model.

4. DESCRIPTION OF THE DATASET

The dataset used in this project is the CelebA dataset. This collection of images is composed of 202,599 faces of celebrities. Each of them is accompanied by 40 binary attributes. These physical attributes are describing each image. The interesting part is the attributes are not equally distributed inside the data. For instance, if the 'Young' characteristic is overly seen in the faces with more than 80% of images linked with this feature, others such as "Bald" or "Chubby" are much rarer. The distribution of each attribute can be seen in the following figure:

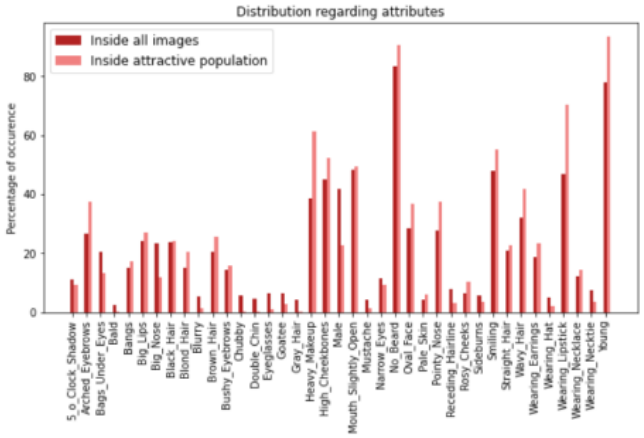

Fig. 1. Distribution inside each attributes regarding attractiveness

These distributions are salutary for us as it is illustrative of real databases. Indeed, it appears often that some of the sensitive attributes are underrepresented and therefore the model is not properly trained on them. We will therefore consider these types of characteristics as protected and study them further. The Bald attribute will be used to test the efficiency of adding data to the training set. We will also explore a threshold adjusting solution for the "Eyeglasses" attribute.
Finally, it can also be interesting to look at the proportion of each attribute inside the "Attractive" population (i.e faces with the attribute "Attractive" True). From this plot, we see how some attributes are decisive when it comes to deciding whether a celebrity is attractive or not. For instance, young women wearing heavy makeup are most likely to be considered attractive whereas old bald people are not. After taking a first look at the dataset, we can now create the classifier itself.

5. CONSTRUCTION OF THE CLASSIFIER

During the first part of the project, we focused on creating an efficient classifier determining if a person is attractive or not based only on their photo. We started by creating a classifier predicting the gender of a person as it was easier for us to evaluate and switched to the attribute of interest after the biggest adjustments and parameterization were completed.
It also seemed to be a good approach since the gender seems to be a deterministic attribute for attractiveness classification, as can be seen in the provided dataset.
If our model easily distinguishes between males and females it will likely be able to get reasonable results to predict attractiveness. This hypothesis was later confirmed when the final attractiveness classifier was constructed.

5.1. Architecture

First, as we are working with images, we chose to use convolution layers for our neural network. This type of layer is indeed really efficient to extract the key elements of a picture and is used in many image classification problems. Then, we used linear layers to output the final prediction.
To evaluate our model and therefore validate our choices of architecture, the CelebA database was split into three subsets. A train subset to train the model, a validation subset used to detect over-fitting and refine our choice of parameters, and finally a test subset to test our network.
After multiple adjustments of the parameters, our final structure is the following :

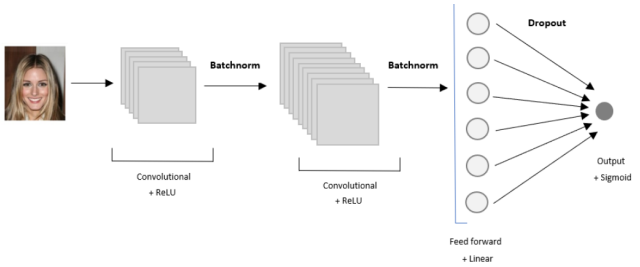

Fig. 2. Architecture of the neural network

Even the architecture is quite simple, it still provided satisfying results in terms of accuracy and speed. The run of the test and validation part, written on a Jupyter notebook, took an average of an hour with GPU on Google Colab.
During this project we focused mostly on the fairness of the model. Thus, having a fast classifier is necessary knowing that we are going to train it multiple times. That is especially true when introducing pre-processing steps that require retraining the model.

Concerning the parameters of the model, we had plenty of leverage points to increase the accuracy of our model and its speed. For instance, the quality of the input image was lowered to gain more efficiency while keeping high performance. In the same order of ideas, the validation process was skipped until we reached a sufficient quality for our classifier. This saved a lot of time in the training while not distorting the exercise as the over-fitting was by all means low.
We also chose to increase the number of channels along the different convolutional layers as it noticeably increased the accuracy.
Another step was determining the size of the batches. We increased the batch size throughout our experiment as it yielded smoother results.
Finally, even though we knew that over-fitting was low in our case, we realized after adding the validation step that it is still quite noticeable. To reduce it, as can be seen on the architecture figure, we added batch normalization as well as dropout.

These two methods helped reduce instability that can appear and spread in the neural network, resulting in an over-fitting behavior.
Indeed, the batch normalization will normalize the entries of each layer and this way avoid having too much disparity. And, by using dropout, we deliberately ignore a proportion of the outputs of the layer. These solutions will promote a regularization amongst the layers.

5.2. Result of our model

The final step is to evaluate the accuracy of our model before trying to assess its fairness. To do so, we searched through literature and compared our results with previous models. In our case, the test accuracy reached 80,3%. A study comparing seven highly performing models
[1] reached an accuracy of 82,52% thus making our outcome very reasonable. The model compared were also using CNNs having many more layers than ours resulting in a more complex and long training. The behavior of the training from our model can be seen below :

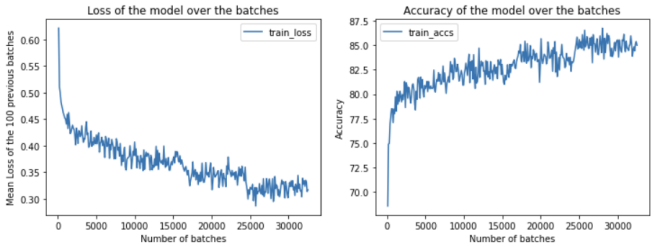

Fig. 3. Training process of the neural network

The over-fitting is not too important and we can see how the train accuracy never stops increasing throughout the epochs. However, as the validation accuracy doesn't improve with the same speed, using only 4 epochs does not affect the result too much.
Finally, it seems relevant to introduce the next part concerning fairness by comparing the distribution of the different attributes inside the initial attractive population versus the predicted attractive population. Finding that some parameters are underrepresented or overrepresented will exhibit some of the flaws of our model. This can be seen in the following figure :

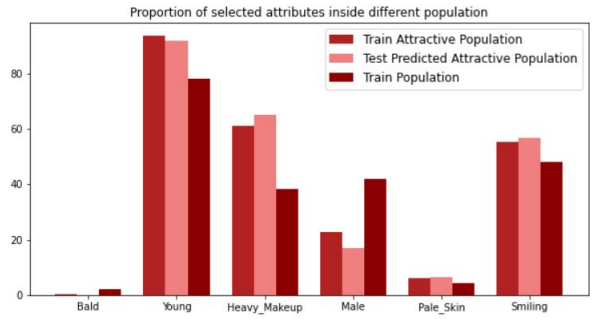

Fig. 4. Distribution of selected attributes regarding the population

The selected attributes from the plot above were chosen because of their distribution in the test dataset. We wanted to compare common attributes as well as rare ones.
Again we can see the disparity between the weight of each attribute when it comes to attractiveness. Young, Heavy Makeup, and Male display big differences between their original proportion in the test data set and their proportion inside the classified attractive population.
We also see that the distribution of each attribute is very similar between the predicted attractive population and the original population. However, the fear of poor accuracy for the underrepresented attributes was justified as our model is almost not predicting any bald people as attractive from the test subset. This is a great concern and the solutions detailed in the following sections will try to overcome this issue.

6. FAIRNESS EVALUATION OF THE CLASSIFIER

6.1. Output of the classifier

By studying the output of the neural network, we can see that some attributes are decisive to be classified as attractive. Several characteristics are very represented amongst the classified attractive and others are almost non-existent.

	% in the training dataset	% predicted by the model
Bald	0.15	0.03
Gray hair	0.18	0.14
Double Chin	0.26	0.42
Chubby	0.31	0.53
Eyeglasses	1.04	0.60

Fig. 5. Worst attributes to have to be classified as attractive

	% in the test dataset	% predicted by the model
Young	92.1	93.14
No beard	92.06	91.77
Wearing lipstick	76.1	78.06
Heavy make-up	63.6	64.88
Smiling	56.49	56.75

Fig. 6. Best attributes to have to be classified as attractive

The statistics shown above are coherent with the images that our classifier considers the most and least attractive.
, according to our model, the epitome of attractiveness is a young person wearing makeup and smiling. On the other hand, being old, having grey hair, wearing eyeglasses, and being chubby makes you unattractive.

Fig. 7. On the left: the image classified as least attractive by our model.On the right: the image classified as most attractive by our model.

6.2. Chosen fairness metrics

Once the classifier built and trained, we wished to evaluate its fairness. The first step was to define the criterion by which we would assess it.
Our choice of fairness metrics was based on the information provided by [2]. At first, we considered the three following formulas :

- Statistical parity: $P(C = 1|A = 0) = P(C = 1|A = 1)$. This metric is also referred to as "Fairness through blindness": in our case, for a given attribute, the probability to be classified as attractive would be the same whether the image possesses that attribute or not. Mathematically this translates into the fact that the proportion of people having the attributes amongst those classified as attractive is equal to the proportion of people having the attribute in the original dataset. However, as underlined by the paper, this metric forces a condition that is not faithful to reality and does not ensure fairness as unqualified individuals can be classified positively to satisfy the metric.
- Equality of opportunity: $P(\widehat{Y} = 1|A = 0, Y = 1) = P(\widehat{Y} = 1|A = 1, Y = 1)$. For a given attribute, this metric focuses on the lack of discrimination for the positive outcome only. This positive outcome is defined depending on the study, in our case, it is being classified as attractive. For example, if you apply for a university, it ensures that whether you are a woman or a man does not matter to get accepted as long as you're qualified. Mathematically this translates to having the same True Positive Rate, $\frac{TruePositive}{TruePositive+FalseNegative}$ for those who have the attribute and those who don't.
- Equalized odds: $P(\widehat{Y} = 1|A = 0, Y = y) = P(\widehat{Y} = 1|A = 1, Y = y), y \in \{0, 1\}$. For a given attribute, this metric enforces the lack of discrimination for both the positive and the negative outcome. For example, if you apply for a university, it ensures that whether you are a woman or a man does not matter to get accepted as long as you're qualified. And if you're not qualified, you're equally probable to be rejected, irrespective of whether you are a man or a woman. Mathematically, this translates to having the same True Positive Rate, $\frac{TruePositive}{TruePositive+FalseNegative}$, and the same False Positive Rate, $\frac{FalsePositive}{FalsePositive+TrueNegative}$, for those who have the attribute and those who don't.

After familiarizing ourselves with these three metrics, we decided to evaluate the fairness of our algorithm by the third one we encountered: Equalized Odds.
The first reason behind this choice is that we wanted a fairness metric that would be faithful to reality, and doesn't erase the fact that some attributes are crucial when it comes to being classified as attractive or not. Thus, we chose not to keep the Statistical Parity measure.
The second reason is that Equalized Odds is a more complete form of the Equality of Opportunity metric. It takes into account the negative and positive outcomes and therefore offers a more complete definition of fairness. As a personal choice, we considered that fairness should not only be achieved when an outcome is favorable but should aim to reflect reality as much as possible.

6.3. Fairness metrics results for our classifier

Even though the fixes we propose in the following sections to make our classifier more fair are based on the Equalized Odds metric, we still wished to measure the fairness of our model with the two other measures presented above.

We chose to focus on the two following attributes: Eyeglasses and Bald. They were one the characteristics that displayed the most disparity between the original and the predicted attractive population. Here are the results of the fairness metrics:

	Bald	Eyeglasses
Statistical parity	0.53 =/= 0.01	0.55 =/= 0.05
True positive rate	0.83 =/= 0.13	0.82 =/= 0.25
False positive rate	0.22 =/= 0.01	0.24 =/= 0.03

Fig. 8. Results of the fairness metrics for the Eyeglasses and the Bald attributes

We can see that none of the fairness metrics are satisfied by our classifier. What's more, the difference is too important to be acceptable. For instance, for the Bald attribute, our model predicts 50 times fewer people than expected to achieve Statistical Parity.
Furthermore, for the Equalized Odds, the metric we wish to satisfy, we can see that the False Positive and True Positive rates are extremely disparate.
It is safe to say at this point that our classifier is unfair and exhibits important biases for these attributes.

Therefore, we decided to try two possible fixes: adjusting the threshold and adding images.

7. FIRST FIX: ADJUSTING THE THRESHOLD

To satisfy the Equalized Odds criterion we first opted for a post-processing tool: adjusting the threshold.
Up to this point, we had set the threshold at 0.5, meaning that if the classifier output a probability superior to 0.5 that the image is attractive, we would classify it as such, and vice-versa. After reading through the [3] paper, we decided to try adjusting the threshold. The aim is to find thresholds for the images having a protected attribute and those that don't so that their respective True Positive Rate and False Positive Rate are equal.
To do so, we used Receiver Operating Characteristic (ROC) curves that are a plot of the True Positive Rate in function of the False Positive Rate. We plot this curve for the pictures having the attributes and those that don't. When the two curves meet, that means that the Equalized Odds metric is satisfied as the True Positive Rate and False Positive Rate of the two groups are equal.
To draw the curves, we launched the test phase for the model with a threshold varying from 0 to 1 by a 0.05 interval and calculated each time the False Positive Rate and the True Positive Rate for both groups.

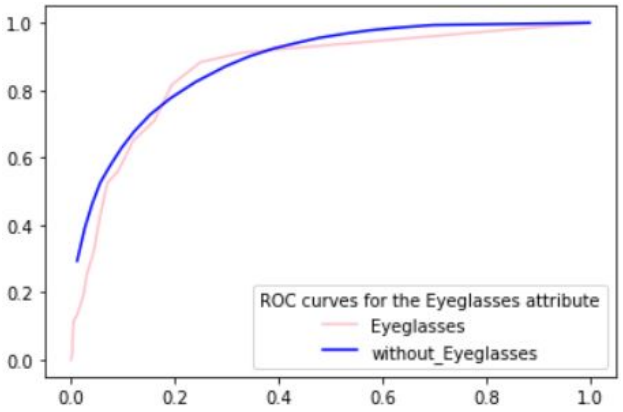

Fig. 9. ROC curves for the Eyeglasses attribute

We then chose the highest point where the curves meet to maximize the True Positive Rate of the classifier, and therefore not sacrifice the accuracy too much. We found the following results:

- **Threshold for image with Eyeglasses:** 0.4
- **Threshold for image without Eyeglasses:** 0.85

Therefore by the thresholds chosen, we're making it, approximately, twice as hard to be classified as attractive when one doesn't have glasses. In order words, the classifier needs to be more than twice as sure to classify someone without glasses as attractive.

Using these thresholds, we obtain the following rate results:

	Eyeglasses (before)	Eyeglasses (after)
True positive rate	0.82 =/= 0.25	0.46 ~= 0.46
False positive rate	0.24 =/= 0.03	0.04 ~= 0.05

Fig. 10. True Positive and False Positive rates before and after threshold adjustment

Thus, we can see that the Equalized Odds metric is approximately satisfied for the Eyeglasses attribute using this threshold adjustment.
Furthermore, the accuracy of the classifier is not immensely affected by this solution: it drops at 79.8% for the 0.4 threshold and 71.3% for the 0.85.

However, this has only been done for one attribute where the curves happened to meet. Other cases might require taking into account many more attributes. For instance, if we want to adjust for multiple protected categories (race, gender etc.), a common threshold needs to be found that maximizes the True Positive Rate for all at the price of increasing the False Positive Rate and therefore deteriorates the network's accuracy.

8. SECOND FIX: BALANCING THE DATASET

The second solution that we tried to improve the fairness of our model is to add information concerning the protected attributes before training our model. Indeed, a cause of fairness deficiency is often the lack of enough data for the network to train on. This is the case for the CelebA dataset concerning the «Bald» people. Indeed, only 3713 photos of bald people are part of the training set. Therefore, we wanted to enrich the data by copying the photos of bald people, perform a small rotation, and adding them. This manipulation multiplied the number of bald people in the dataset by ten, reaching 37130 photos.

By comparing the model trained with the new data with our original trained model, we could see a little improvement. However, the percentage of occurrence of bald people inside the attractive population did not change. The results are not outstanding. It is normal as the new photos added to the original data-set are not carrying more visual information to our model. They are still showing the same faces and therefore the model is more likely to detect specific pictures rather than learning to deal with bald people in general. Still, adding more pictures forces the model to understand that it needs to recognize these photos and that explains why the results are a bit better. The detailed results are shown below :

	% of occurrence inside attractive population	% of occurrence in predicted attractive population	
		Without data added	With data added
Bald	0,15	0,03	0,03
Gray_Hair	0,18	0,14	0,11
Double_Chin	0,26	0,42	0,51
Chubby	0,31	0,53	0,56
Eyeglasses	1,04	0,6	0,58

Fig. 11. Percentage of attributes inside predicted attractive population between the two methods

We see that we get the same number of bald people predicted as attractive with our method compared to the original data set. Some variables had a different but the method still is inefficient and unreliable. Therefore in the future, it would be better to introduce new unique data allowing the model to learn better by observing new information and handle the protected attribute.

9. CONCLUSION

In conclusion, after having built and trained our model we were able to detect some clear biases exhibited towards certain categories.
To quantify this lack of fairness, literature provided us with multiple metrics allowing us to choose what we wanted to focus on: partiality for both positive and negative outcome.
Based on this formula, we explored solutions to try to alleviate the biases that were more or less effective.
Threshold adjustment worked fine on a small example for a particular attribute, but is a questionable method at a larger scale that requires long calculations.
Adding more data into the training dataset seemed to be the most relevant fix. By showing the model more diversity and insisting upon each, it would have allowed it not to neglect certain categories with protected attributes during testing. However, by only adding existing data and twisting a bit, we didn't achieve any clear improvement.
Still, we think that in the general case, where a classifier is used for more life-changing choices such as loans and recruitment, the best solution is to have a balanced dataset that is faithful to reality.
Through this fun example, we saw how big of a bias a neural network can introduce through classification and we became more conscious of the effort that needs to be made to have a fairer artificial intelligence. It starts at the very beginning: collecting data that is inclusive and representative of reality.

10. REFERENCES

[1] Youngkyoon Jang, Hatice Gunes, and Ioannis Patras, "Registration-free face-ssd: Single shot analysis of smiles, facial attributes, and affect in the wild," *Computer Vision and Image Understanding*, vol. 182, 02 2019.

[2] Moritz Hardt, Eric Price, and Nathan Srebro, "Equality of opportunity in supervised learning," 2016.

[3] Elias Baumann and Josef Lorenz Rumberger, "State of the art in fair ml: From moral philosophy and legislation to fair classifiers," 2018.