Algorithmic fairness

- Is our CelebA smile predictor upholding the standards of fairness?

Charlotte Friis Theisen
s143922@student.dtu.dk
{'out': tensor([-2.7959, 2.7955])}

Aleksander Pratt
s153642@student.dtu.dk
{'out': tensor([1.2387, -1.2403])}

Martin Johnsen
s144731@student.dtu.dk
{'out': tensor([-1.7870, 1.7862])}

Motivation

Deep Learning algorithms are increasingly used to affect people's lives. Therefore, these algorithms must be fair and unbiased. The emerging field of algorithmic fairness investigates such issues by providing defined metrics to measure if an algorithm is discriminating or not.
This project seeks to investigate whether a social bias on *Smiling/Not Smiling*-prediction can be found for different demographic groups.

Results

= bias detected

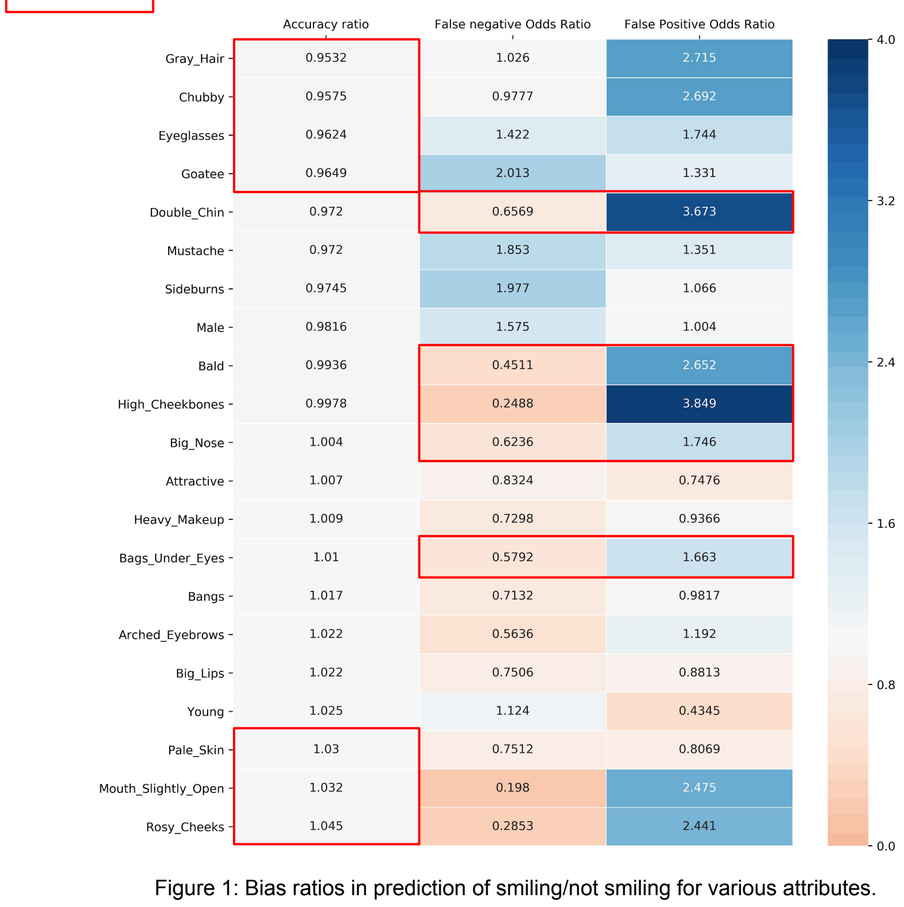


Figure 1: Bias ratios in prediction of smiling/not smiling for various attributes.

Fairness metric

The fairness metric used to evaluate the models is given from the *Equalised Odds*-definition:

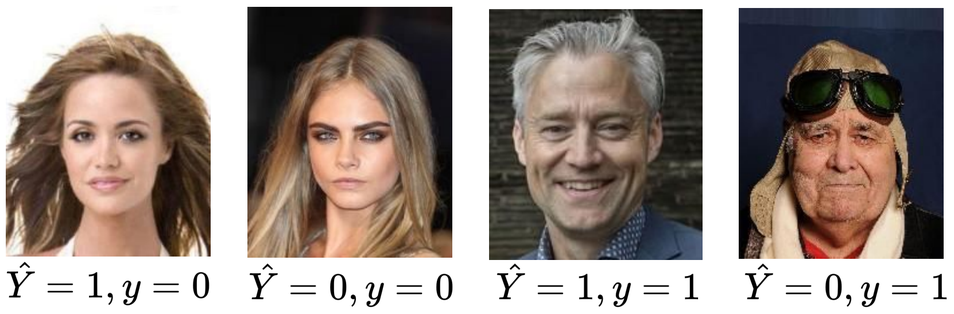
$$Pr\{\hat{Y}|A=0,Y=y\} = Pr\{\hat{Y}=1|A=1,Y=y\}, y \in \{0,1\}$$

The definition states that the predictor variable, $\hat{Y}$, should satisfy equalised odds with respect to a protected attribute, A, and target Y, if $\hat{Y}$ and A are independent conditional on Y.
The definition is relaxed by some factors to set a threshold for when a bias is identified:

$$Acc_{Ratio} = \frac{Acc_{A=1}}{Acc_{A=0}} \geq 1.3 \;\; OR \;\; \frac{Acc_{A=1}}{Acc_{A=0}} \leq 0.97$$

$$FN_{OddsRatio} = \frac{FN_{A=1}}{FN_{A=0}} \geq 1.5 \;\; AND \;\; FP_{OddsRatio} = \frac{FP_{A=1}}{FP_{A=0}} \leq 0.75$$

Are these people smiling?



$\hat{Y}=1, y=0$ $\hat{Y}=0, y=0$ $\hat{Y}=1, y=1$ $\hat{Y}=0, y=1$

Model architecture

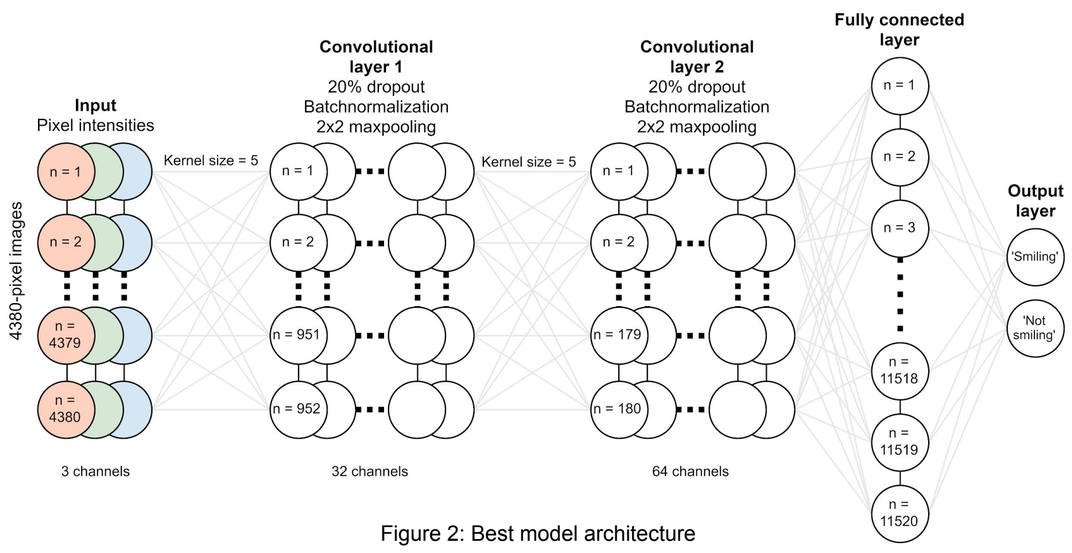


Figure 2: Best model architecture

Of 18 evaluated models, the best one is an architecture of 2 convolutional layers with 32 and 64 channels respectively (model 2 in table 2). The model uses a ReLu activation function, a dropout rate of 0.2, batchnorm and maxpooling in both layers. The associated training and validation accuracy for the best model is plotted in figure 3. The accuracy in predicting whether a person is smiling or not is 91.57%. In comparison, similar research on identical data obtains a smile accuracy of 89.34% using various cropping schemes.

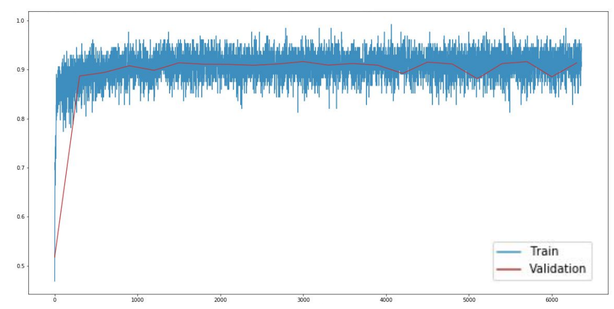


Figure 3: Training and validation accuracy for the best model

Dataset

The CelebFaces Attributes comprises a large dataset of celebrities' faces in an aligned format. The images hold different poses and color variations and the dataset has a total of:

- 10,177 identities
- 202,599 aligned face images
- 40 binary attributes

Further work

The current findings in this project, reveal a bias in several attributes, e.g. Chubby, Goatee, and Bald, as indicated in the results section. Further work will be aimed at handling these biases via approaches such as upsampling or a weighted loss function. The goal is to eliminate biases that might entail social biases in future AI systems and computer vision tasks.

Model tuning

Tuning has been carried out on GPU via Amazon Web Services (AWS). Initially, 10 models with different tuning parameters were assessed from their smile predicting validation accuracy (Table 1). The conclusions were used to design eight complex models that were tested in run 2 (Table 2). All models are trained for 5 epochs, with batch size 128, using Adam optimizer with default learning rate, and a stride of 1.

Model	Layers	Channels	Activation	Accuracy
1	2 x CNN	32, 64	relu, tanh	**0.9108**
2	2 x CNN	64, 128	relu, tanh	0.9073
3	2 x CNN	32, 64	2xrelu	0.9067
4	2 x CNN	64, 128	2xrelu	0.9078
5	1 x CNN	32	tanh	0.8858
6	1 x CNN	128	tanh	0.8967
7	2 x CNN	32, 64	tanh, relu	0.8947
8	2 x CNN	64, 128	tanh, relu	0.8984
9	2 x CNN	32, 64	2xtanh	0.8810
10	2 x CNN	64, 128	2xtanh	0.8746

Table 1: Initial model tuning

Model	Layers	Channels	Activation	Dropout	Maxpool	Batchnorm	Weight decay	Accuracy
1	2 x CNN	32, 64	relu, tanh	0.2	2x2	TRUE	0.01	0.9105
2	2 x CNN	32, 64	2xrelu	0.2	2x2	TRUE	0.01	**0.9157**
3	2 x CNN	64.128	2xrelu	0.2	2x2	TRUE	0.01	0.8902
4	2 x CNN	32, 64	2xrelu	0.8	2x2	TRUE	0.1	0.9136
5	3 x CNN	32,64,128	2xrelu, tanh	0.2	2x2	TRUE	0.01	0.9146
6	4 x CNN	16,32,64,128	3xrelu, tanh	0.2	2x2	TRUE	0.01	0.9061
7	3 x CNN	32,64,128	3xrelu	0.2	2x2	TRUE	0.01	0.9123
8	4 x CNN	16,32,64,128	3xrelu, tanh	0.2	2x2	TRUE	0.01	0.9119

Table 2: Further model tuning elaborating on results from table 1