02456 - Deep Learning
Victor Célérier - s201836
Zineb Fadili - s201501

Fairness in Computer Vision

Fairness in Machine Learning

Neural networks are making choices in many fields
Finance, insurance, social media.. and many more

Looking into fairness is now urgent

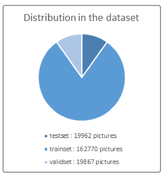
Which motivates our project

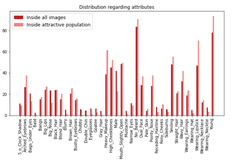
Building a classifier for the attractive attribute

Identifying Bias

Alleviating them

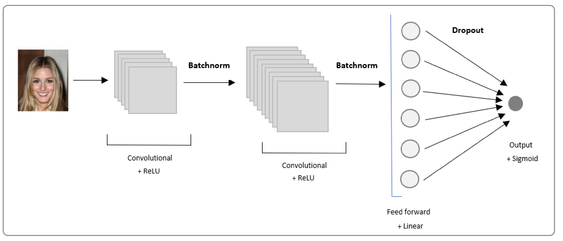
The CelebA Dataset





Our Model

Architecture



Training Accuracy:
81.3%

Validation Accuracy:
78.8%

Test Accuracy:
80.3%

Steps taken

Batch Size

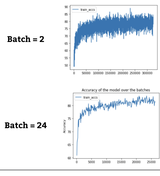


Image Size

Reducing Overfitting

Batch norm + Dropout

Fairness of Our Model

What decides attractiveness?

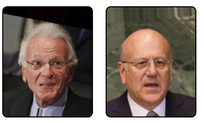
Top attributes to have

	% 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

Two persons predicted as most atttractive

Worst attributes to have

	% 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



Two persons predicted as least atttractive

Evaluating the model

Fairness metrics

Statistical parity - Fairness through blindness

$$\mathbb{P}\{C = 1 \mid A = 0\} = \mathbb{P}\{C = 1 \mid A = 1\}$$

Equality of opportunity - Equal favorable outcome (TPR)

$$\Pr\{\hat{Y} = 1 \mid A = 0, Y = 1\} = \Pr\{\hat{Y} = 1 \mid A = 1, Y = 1\}$$

Equalized odds - Equal TPR and FPR

$$\Pr\left\{\hat{Y} = 1 \mid A = 0, Y = y\right\} = \Pr\left\{\hat{Y} = 1 \mid A = 1, Y = y\right\}, \quad y \in \{0, 1\}$$

Fairness results

Metrics values for two attributes

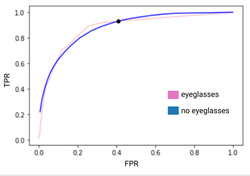
	Bald	Eyeglasses
Statistical parity	0.53 =/= 0.01	0.55 =/= 0.05
True positive rate	0.83 =/= 0.13	0.82 =/= 0.25
False positive rate	0.22 =/= 0.01	0.24 =/= 0.03

The model is not fair..

Improvements needed to achieve equalized odds

First improvement : treshold adjustment

Roc curve for the 'Eyeglasses' attribute



Tresholds chosen :

Without eyeglasses: 0.8

With eyeglasses: 0.4

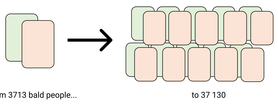
New results:

	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

But this causes a loss in accuracy...

Second improvement : equilibrating the dataset

Goal : Giving more weight to bald peoples



From 3713 bald people... to 37 130

Results

The model did not learn more, we need new images, not copy

	% 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

	% of occurrence inside attractive population	% of occurrence in predicted attractive population	
		Without data added	With data added
Young	92,1	91,77	91,73
No_Beard	92,06	93,14	93,75
Wearing_Lipstick	76,1	78,06	78,88
Heavy_Makeup	63,6	64,88	65,65
Smiling	56,49	56,75	57,19

The best way to tackle fairness issue is to gather new data