Fairness is in the details : Face Dataset Auditing

Key ideas

V. Lafargue[1,2,3] E. Claeys [4] J.-M. Loubes [2,3]

1: IMT : Institut de Mathématiques de Toulouse

2: INRIA : Institut national de recherche en sciences et technologies du numérique

3: ANITI 2 : Artificial and Natural Intelligence Toulouse Institute 2

4: IRIT : Institut de Recherche en Informatique de Toulouse

ECML PKDD 2025

Introduction

- What is an audit? Analyzing decision-making, sensitive variables such as Gender (G), Age, Ethnicity with Fitzpatrick scale (Fitz)

Introduction

- What is an audit? Analyzing decision-making, sensitive variables such as Gender (G), Age, Ethnicity with Fitzpatrick scale (Fitz)
- Why audit an image dataset & Difficulty? Image inherently encodes explicit features

Introduction

- What is an audit? Analyzing decision-making, sensitive variables such as Gender (G), Age, Ethnicity with Fitzpatrick scale (Fitz)
- Why audit an image dataset & Difficulty? Image inherently encodes explicit features
- Bias definition : distribution equivalence for:
  1. Parity test (one sensitive variable):
     $\mathcal{H}_0 : \mathcal{D}(\{G\}) \sim B(p)$, expected distribution (i.e., gender : $p = \frac{1}{2}$)
  2. Equal representation test (two sensitive variables):
     $\mathcal{H}_0 : \mathcal{D}(\{\text{Fitz}|G = \text{Male}\}) \sim \mathcal{D}(\{\text{Fitz}|G = \text{Female}\})$

Introduction

- What is an audit? Analyzing decision-making, sensitive variables such as Gender (G), Age, Ethnicity with Fitzpatrick scale (Fitz)
- Why audit an image dataset & Difficulty? Image inherently encodes explicit features
- Bias definition : distribution equivalence for:
  1. Parity test (one sensitive variable): $\mathcal{H}_0 : \mathcal{D}(\{G\}) \sim B(p)$, expected distribution (i.e., gender : $p = \frac{1}{2}$)
  2. Equal representation test (two sensitive variables): $\mathcal{H}_0 : \mathcal{D}(\{\text{Fitz}|G = \text{Male}\}) \sim \mathcal{D}(\{\text{Fitz}|G = \text{Female}\})$
- Ensuring result : statistical guarantees
  1. χ^2 test
  2. CLT-based test
  3. Wasserstein-based test

Table of Contents

Table of Contents

Neural Network Architecture, with pixel information feature extraction

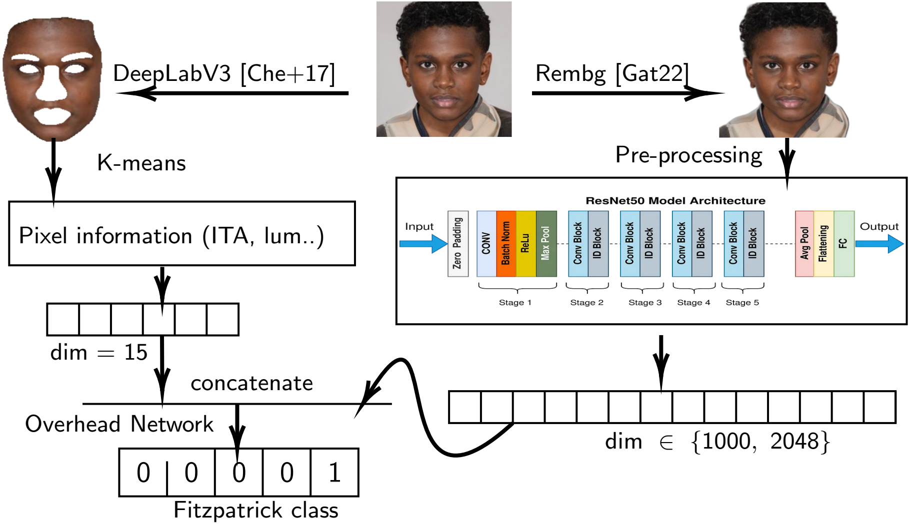
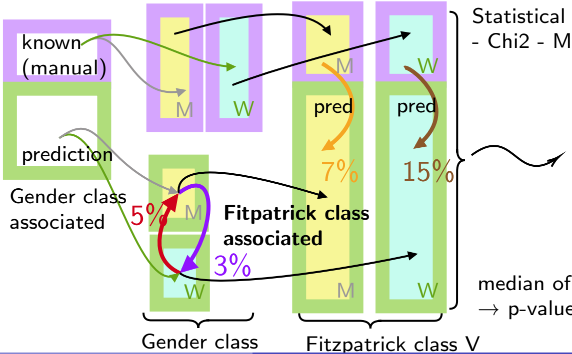


Diagram explaining our error-aware testing pipeline in an equal representation test of the Fitzpatrick class V conditioned by the gender

Accuracy / Precision (%)		
Gender (G)	Men (M) : 97 % →3% permutation with W	Women (W) : 95 % → 5% permutation with M
Fitzpatrick class V given G ($Fitz._V$)	V knowing predicted M : 93 % $\{\widehat{X}_{Fitz._V} \| \widehat{X}_G = M\} \leftarrow$ 7% replacement $x \sim \mathcal{U}(X_{Fitz._V} \| X_G = M, X_{pred} = False)$	V knowing predicted W : 85 % $\{\widehat{X}_{Fitz._V} \| \widehat{X}_G = W\} \leftarrow$ 15% replacement $x \sim \mathcal{U}(X_{Fitz._V} \| X_G = W, X_{pred} = False)$



Statistical test:
- Chi2 - Mean test - Wasserstein test

median of test results (one per simulation)
→ p-value < threshold

Table of Contents

1. Generated images[1] by GAN: 10k imgs (aims for ethnic diversity)
2. Subset of CelebA[2]: 1.5K real images

Figure: Fitzpatrick classification (5 class) from the left to right Phototype I, II, III-IV, V, VI. The first row is GAN dataset, the second is the CelebA dataset.

[1]https://generated.photos/solutions/academic-research
[2]Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV (December 2015)

Age distribution	0-2	3-9	10-19	20-29	30-39	40-49	50-59	60-69	70+
Real [wor22][our22]	2	12	17	17	15	13	11	8	7
Predicted CelebA	0	0.3	3	48	26	11	7	4	0.1
Predicted GAN	0.3	4	6	61	22	4	2	0.3	0.05

Table: Age distribution : Real estimated and our datasets (%)

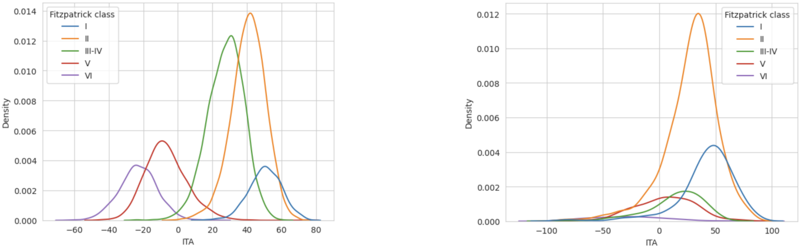


Figure: Probability density of Individual Typology Angle (ITA) given the Fitzpatrick class, GAN distribution on the left, CelebA distribution on the right. ITA implementation by [TJX23; Mer+19].

Table of Contents

Architecture and hyperparameters usefulness

Hyperparameter	GAN	CelebA
Skin information (ITA..)	+0.6	+0.2
Overhead network choice	+3.7	+4.2
Skin mask	-1.7	+1.2
Removing background	+0.1	+1.8
Latent space size (2048)	+1.9	+2.7

Table: Impact of hyperparameters and architecture for neural network designed to classify the Fitzpatrick scale on the GAN and CelebA datasets.

Parity results (and ablation)

Sensitive Variable	Sample Size				
	100	500	1000	3000	8000
Gender	×	×	×	×	×
Age					
0-2	✓	✓	✓	✓	✓
3-9	×	×	×	×	×
10-19	×	×	×	×	×
20-29	×	×	×	×	×
30-39	×	×	×	×	×
40-49	×	×	×	×	×
50-59	×	×	×	×	×
60-69	×	×	×	×	×
70+	×	×	×	×	×
Fitzp. Class					
I	×	×	×	×	×
II	×	×	×	×	×
III- IV	×	×	×	×	×
V	×	×	×	×	✓
VI	×	×	×	×	×

(a) GAN

Sensitive Variable	Sample Size			
	100	500	1000	1200
Gender	×	×	×	×
Age				
0-2	✓	✓	✓	✓
3-9	×	×	×	×
10-19	×	×	×	×
20-29	×	×	×	×
30-39	×	×	×	×
40-49	✓	✓	✓	✓
50-59	✓	✓	✓	✓
60-69	×	×	×	×
70+	×	×	×	×
Fitzp. Class				
I	×	×	×	×
II	×	×	×	×
III- IV	×	×	×	×
V	×	×	×	✓
VI	×	×	×	$\checkmark_{2/3}$

(b) CelebA

Statistical test on the parity of Gender ($\mathcal{H}_0 : p = 0.5$) and the real distribution for the Fitzpatrick class and Age ($\mathcal{H}_0$: marginal and global distributions are equivalent). ✓, $\checkmark_{2/3}$ and × respectively means that 0, 1 or at least 2 tests rejected $\mathcal{H}_0$. Colored cell means that the test result are different because of the error-aware protocol: ■ means that, thanks to the error-aware method, less test rejected $\mathcal{H}_0$.

Equal representation results (and ablation)

Sensitive Variable	Sample Size 100	500	1000	3000	8000
Age					
0-2	✓	✓	✓	✓	✓
3-9	✓	✓	×	✓	✓
10-19	✓	✓	✓	✓	✓
20-29	✓ 2/3	✓	×	✓	✓ 2/3
30-39	✓	✓	✓	✓	✓
40-49	✓	✓	✓	✓	✓
50-59	✓	✓	✓	✓	✓
60-69	✓	✓	✓	✓	✓
70+	✓	✓	✓	✓	✓
Fitz. Class					
I	✓ 2/3	✓ 2/3	×	×	×
II	×	×	✓ 2/3	✓ 2/3	✓
III- IV	×	×	×	×	✓ 2/3
V	×	×	✓	×	✓ 2/3
VI	×	×	×	×	×

(a) GAN

Sensitive Variable	Sample Size 100	500	1000	1200
Age				
0-2	✓	✓	✓	✓
3-9	✓	✓	✓	✓
10-19	✓	✓	✓	✓
20-29	×	×	×	×
30-39	✓	✓ 2/3	✓	✓
40-49	✓	✓	✓	✓
50-59	✓	✓	✓	✓
60-69	✓	✓	✓	✓
70+	✓	✓	✓	✓
Fitz. Class				
I	✓ 2/3	✓ 2/3	×	×
II	×	×	×	×
III- IV	✓	✓	✓	✓
V	×	×	✓ 2/3	✓
VI	✓	×	✓ 2/3	✓ 2/3

(b) CelebA

Equal repr. test on the for the Fitz. class and the Age, with respect to each reflected closest binary gender subgroup. $\mathcal{H}_0$: The Fitz. or Age distribution of the men subgroup is the same as the women subgroup. and respectively mean that, because of the error-aware method, more test or respectively less test rejected $\mathcal{H}_0$.

Table of Contents

References

[Che+17] Liang-Chieh Chen et al. *Rethinking Atrous Convolution for Semantic Image Segmentation*. 2017.

[Gat22] Daniel Gatis. *Rembg*. 2022. URL: https://github.com/danielgatis/rembg.

[Mer+19] Michele Merler et al. *Diversity in Faces*. 2019.

[our22] ourworldata. *Our World in Data*. 2022. URL: https://ourworldindata.org/.

[TJX23] William Thong, Przemyslaw Joniak, and Alice Xiang. *Beyond Skin Tone: A Multidimensional Measure of Apparent Skin Color*. 2023. arXiv: 2309.05148 [cs.CV].

[wor22] worldbankdata. *World Bank Data*. 2022. URL: https://data.worldbank.org/.

Conclusion

- New bias auditing method with the following properties :
  1. minimize the need for manual annotation (100-1000 annotations)
  2. robust to errors in automated annotation (tools against false rejection of $\mathcal{H}_0$)
- Highlighted discrepancies in tests results $\rightarrow$ Majority vote
- For applied cases : Importance of assessing the representativity of individuals
  1. training dataset
  2. when analyzing the results

.