Mitigating Gender Bias in Face Recognition (ICML 2022)

J.R. Conti[⋆†], N. Noiry[⋆], V. Despiegel[†], S. Gentric[†] and S. Clémençon[⋆]
⋆ TELECOM Paris, † IDEMIA

Motivation

Face Recognition systems have been shown to exhibit **biases**, depending on the gender, skin color, age of the person being recognized. Contrary to a common thought, balanced datasets **are not enough** to mitigate gender bias in Face Recognition (see [1]).

Face Recognition (Verification)

Goal: decide whether two face images (not in the training set) correspond to the same identity or not.

How ? By learning an encoder function $f_\theta : \mathbb{R}^{h\times w\times c} \to \mathbb{R}^d$ that embeds the images in a way to bring same identities closer together. The encoder is usually a CNN, trained as an identity classification task.

The closeness between two embeddings $\boldsymbol{z}_i = f_\theta(\boldsymbol{x_i}), \boldsymbol{z}_j = f_\theta(\boldsymbol{x_j})$ is usually quantified with the cosine similarity measure $s(\boldsymbol{z}_i, \boldsymbol{z}_j) := \boldsymbol{z}_i^\intercal \boldsymbol{z}_j / (||\boldsymbol{z}_i|| \cdot ||\boldsymbol{z}_j||)$. This has led the community to normalize the embeddings $\boldsymbol{z}_i$ during training so that the new embeddings lie within a hyperspere. An operating point $t \in [-1, 1]$ has to be chosen to classify a pair $(\boldsymbol{z}_i, \boldsymbol{z}_j)$ as *genuine* (same identity) if $s \geq t$ and *impostor* (distinct identities) otherwise.

Evaluation metrics. For a pair of images having embeddings Z_1, Z_2 and identity labels y_1, y_2:

$$\mathrm{FAR}(t) := \mathbb{P}(s(Z_1, Z_2) \geq t \mid y_1 \neq y_2)$$

$$\mathrm{FRR}(t) := \mathbb{P}(s(Z_1, Z_2) < t \mid y_1 = y_2).$$

The canonical evaluation metric is:

$$\mathrm{FRR@}(\mathrm{FAR} = \alpha) := \mathrm{FRR}(t) \text{ with } \mathrm{FAR}(t) = \alpha.$$

Fairness in Face Recognition

To incorporate fairness with respect to a given discrete sensitive attribute that can take $A > 1$ different values, we consider intra-group metrics, defined for each subgroup $a \in \{0, 1, \ldots, A-1\}$ as:

$$\mathrm{FAR}_a(t) = \mathbb{P}(s(Z_1, Z_2) \geq t \mid y_1 \neq y_2,\ a_1 = a_2 = a)$$

$$\mathrm{FRR}_a(t) = \mathbb{P}(s(Z_1, Z_2) < t \mid y_1 = y_2,\ a_1 = a_2 = a)$$

where a_1, a_2 are the attributes of the pair of face images. Evaluating fairness by a point-wise comparison of the intra-group ROC curves $\mathrm{FRR}_a\mathrm{@}(\mathrm{FAR}_a = \alpha)$ is not a good idea (see [2]). We introduce two new fairness metrics:

$$\mathrm{BFRR}(\alpha) := \frac{\max_{a\in\{0,1\}} \mathrm{FRR}_a(t)}{\min_{a\in\{0,1\}} \mathrm{FRR}_a(t)}$$

$$\mathrm{BFAR}(\alpha) := \frac{\max_{a\in\{0,1\}} \mathrm{FAR}_a(t)}{\min_{a\in\{0,1\}} \mathrm{FAR}_a(t)},$$

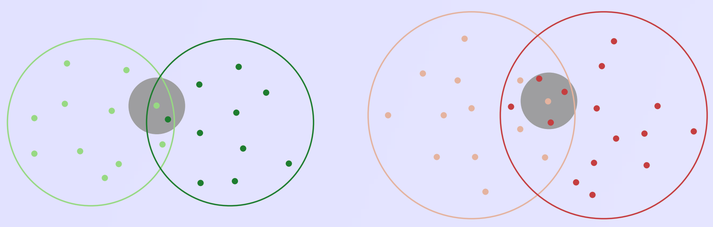
where t is taken s.t. $\max_{a\in\{0,1\}} \mathrm{FAR}_a(t) = \alpha$.

References

[1] Albiero et al. *How Does Gender Balance In Training Data Affect FR Accuracy ?* IJCB, 2020.
[2] Krishnapriya et al. *Issues related to face recognition accuracy varying based on race and skin tone*? IEEE Transactions on Technology and Society, 2020.
[3] Grother et al. *Ongoing face recognition vendor test (frvt) part 3: Demographic effects*? NIST, 2019.
[4] Dhar et al. *PASS: Protected attribute suppression system for mitigating bias in FR*, ICCV, 2021.

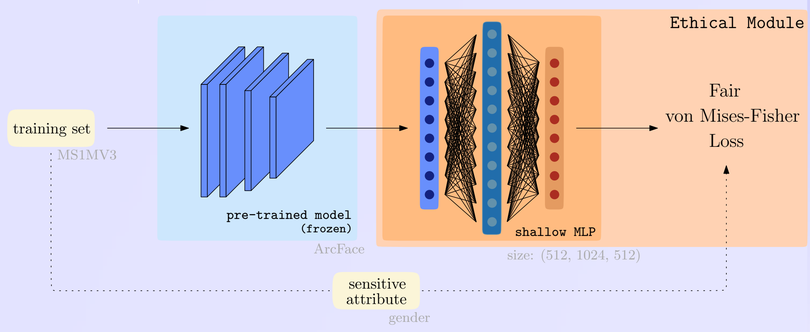
Geometric View on Bias

Females are at a disadvantage compared to males in terms of both FAR and FRR (see [3]). Typically, this is due to *(i)* a smaller repulsion between female identities and/or *(ii)* a greater intra-class variance (spread of embeddings of each identity) for female identities. The next figure illustrates the geometric nature of bias. Each point is the embedding of an image. In green: two male identities. In red: two female identities. The overlapping region between two identities is higher for females than for males. The grey circles are the acceptance zones, centered around an embedding of reference, associated to a constant threshold t of acceptance.



Ethical Module

The Ethical Module is a post-processing method to debias the embeddings output by a frozen pre-trained model. Besides a simple architecture and a fast training (few hours), the Ethical Module enjoys several benefits such as taking advantage of foundation models and the fact that no sensitive attribute is used during deployment.



Fair von Mises-Fisher Loss

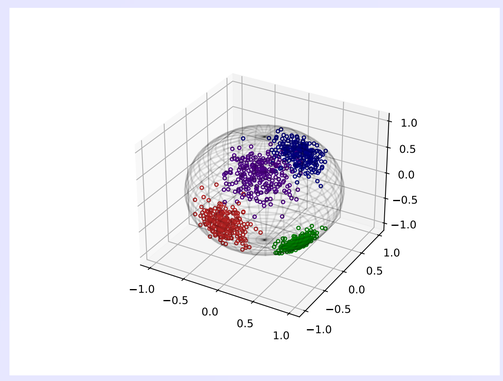
Each identity is statistically modelled as a gaussian conditioned to live within the face hypersphere.

The von Mises-Fisher distribution. The vMF distribution in dimension d with concentration parameter $\kappa > 0$ and mean direction $\boldsymbol{\mu} \in \mathbb{S}^{d-1}$ is a probability measure defined on the hypersphere $\mathbb{S}^{d-1}$ by:

$$V_d(\boldsymbol{z}; \boldsymbol{\mu}, \kappa) := C_d(\kappa) e^{\kappa \boldsymbol{\mu}^\intercal \boldsymbol{z}},$$

with $C_d(\kappa)$ being the renormalization constant.

Mixture model. The model is extended to include all the K identities from the training set by considering a mixture model where each component k $(1 \leq k \leq K)$ is equiprobable and follows a vMF distribution $V_d(\boldsymbol{z}; \boldsymbol{\mu}_k, \kappa_{a_k})$. The next figure displays a vMF mixture model in dimension 3 with 4 components.

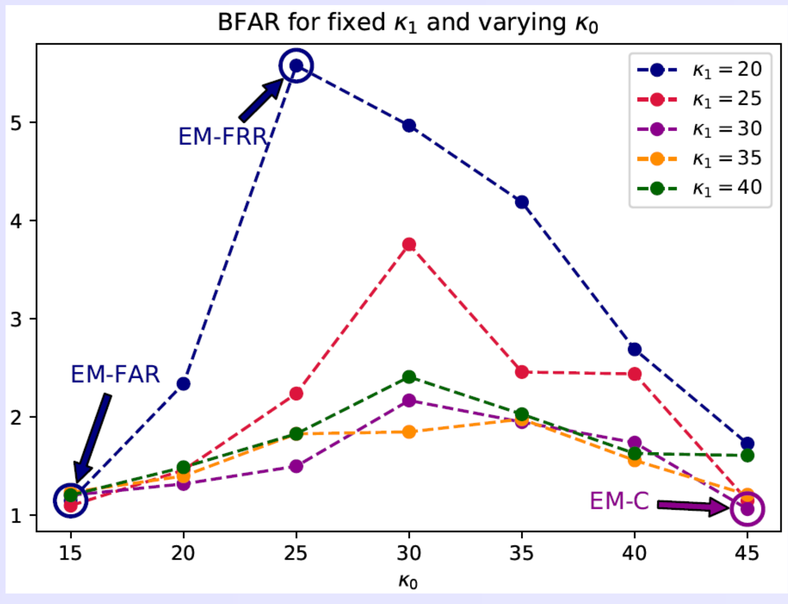


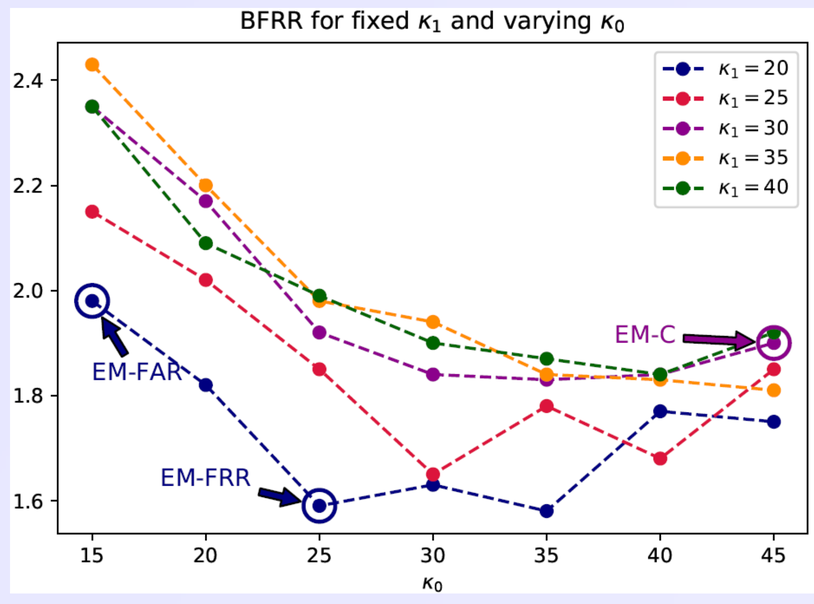
Maximizing the log-likelihood of the model amounts to minimizing the Fair-vMF loss:

$$\mathcal{L}_{\text{FvMF}} = -\frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{C_d(\kappa_{a_{y_i}}) e^{\kappa_{a_{y_i}} \boldsymbol{\mu}_{y_i}^\intercal \boldsymbol{z}_i}}{\sum_{k=1}^{K} C_d(\kappa_{a_k}) e^{\kappa_{a_k} \boldsymbol{\mu}_k^\intercal \boldsymbol{z}_i}} \right),$$

with κ_0, κ_1 taken as hyperparameters.

Grid-Search on IJB-C Dataset





Evaluation on LFW Dataset

With the pre-trained model ArcFace, we compare the Ethical Module methodology to the current state-of-the-art post-processing method for gender bias in Face Recognition: PASS-g (see [4]).

FAR LEVEL:	10^{-4}			10^{-3}		
MODEL	FRR@FAR (%)	BFRR	BFAR	FRR@FAR (%)	BFRR	BFAR
ARCFACE	**0.078**	10.27	4.72	<u>0.059</u>	<u>4.17</u>	1.81
ARCFACE + PASS-G	0.315	**4.54**	6.51	0.107	5.22	2.11
ARCFACE + EM-FAR	0.151	11.22	**2.11**	0.072	9.16	**1.19**
ARCFACE + EM-FRR	<u>0.100</u>	<u>5.89</u>	33.65	**0.058**	**4.11**	5.24
ARCFACE + EM-C	0.164	9.18	<u>2.44</u>	0.081	5.15	<u>1.20</u>

Acknowledgements

This research was partially supported by the French National Research Agency (ANR), under grant ANR-20-CE23-0028 (LIMPID project).