Mitigating Gender Bias in Face Recognition using the von Mises-Fisher Mixture Model

ICML 2022

Jean-Rémy Conti[⋆1,2], Nathan Noiry[⋆1], Vincent Despiegel[2], Stéphane Gentric[2], Stéphan Clémençon[1]

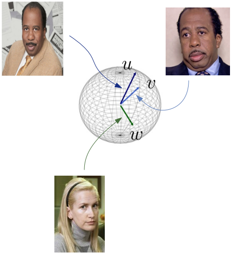
[⋆]Equal contribution

[1]Télécom Paris, [2]IDEMIA

Face Recognition (Verification)

Face Recognition systems use face embeddings which are normalized (they lie on the hypersphere $\mathbb{S}^{d-1}$).

The similarity between two faces is usually measured by the cosine similarity.



Decision rule : $t \in [-1, 1]$, fixed threshold.

- $\langle u, v \rangle \geq t \Rightarrow$ "same identity",
- $\langle u, w \rangle < t \Rightarrow$ "distinct identities".

Evaluation Metric

Two kinds of errors:

- False Positives : predicting "same identity" for two faces from distinct identities. $\rightsquigarrow$ False Acceptance Rate: $\mathrm{FAR}(t)$.
- False Negatives : predicting "distinct identities" for two faces from a same identity. $\rightsquigarrow$ False Rejection Rate: $\mathrm{FRR}(t)$.

Evaluation Metric

Two kinds of errors:

- False Positives : predicting "same identity" for two faces from distinct identities. $\rightsquigarrow$ False Acceptance Rate: $\mathrm{FAR}(t)$.
- False Negatives : predicting "distinct identities" for two faces from a same identity. $\rightsquigarrow$ False Rejection Rate: $\mathrm{FRR}(t)$.

In practice :

1. A threshold $t \in [-1, 1]$ is set to get a deemed acceptable security level α for $\mathrm{FAR}(t)$.
2. The False Rejection Rate is computed at this threshold:

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

Typically $\alpha = 10^{-1}, 10^{-2}, \ldots, 10^{-8}$.

How to Measure Fairness ?

Context

Some algorithms make 10 times more errors on black women than on white men[1].

- $\mathcal{G}$: set of subgroups of the population.
  Examples : women, men, young, old ...
- For all $g \in \mathcal{G}$, we can compute $\mathrm{FAR}_g(t)$ and $\mathrm{FRR}_g(t)$, the False Acceptance and False Rejection Rates, specific to subgroup g.

[1] Grother et al. *Ongoing face recognition vendor test (frvt) part 3: Demographic effects*? NIST, 2019.

How to Measure Fairness ?

Context

- $\mathcal{G}$: set of subgroups of the population.
- For all $g \in \mathcal{G}$, we can compute $\mathrm{FAR}_g(t)$ and $\mathrm{FRR}_g(t)$, the False Acceptance and False Rejection Rates, specific to subgroup g.

Our new fairness metrics

1. Two ratios $\rightsquigarrow$ interpretable metrics:

$$\frac{\max_g \mathrm{FAR}_g(t)}{\min_g \mathrm{FAR}_g(t)} \qquad \text{and} \qquad \frac{\max_g \mathrm{FRR}_g(t)}{\min_g \mathrm{FRR}_g(t)}$$

[1] Grother et al. *Ongoing face recognition vendor test (frvt) part 3: Demographic effects*? NIST, 2019.

How to Measure Fairness ?

Context

- $\mathcal{G}$: set of subgroups of the population.
- For all $g \in \mathcal{G}$, we can compute $\mathrm{FAR}_g(t)$ and $\mathrm{FRR}_g(t)$, the False Acceptance and False Rejection Rates, specific to subgroup g.

Our new fairness metrics

1. Two ratios $\rightsquigarrow$ interpretable metrics:

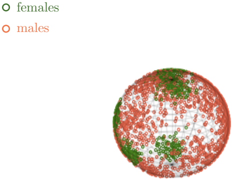
$$\mathrm{BFAR}(\alpha) = \frac{\max_g \mathrm{FAR}_g(t)}{\min_g \mathrm{FAR}_g(t)} \quad \text{and} \quad \mathrm{BFRR}(\alpha) = \frac{\max_g \mathrm{FRR}_g(t)}{\min_g \mathrm{FRR}_g(t)}$$

2. The threshold t satisfies $\max_{g \in \mathcal{G}} \mathrm{FAR}_g(t) = \alpha$ instead of $\mathrm{FAR}_{\mathrm{total}}(t) = \alpha$. $\rightsquigarrow$ more robust to a change of evaluation dataset

[1] Grother et al. *Ongoing face recognition vendor test (frvt) part 3: Demographic effects*? NIST, 2019.

Geometric Embedding View on Fairness

Observation : The embeddings of women fill less space on the hypersphere than the embeddings of men.



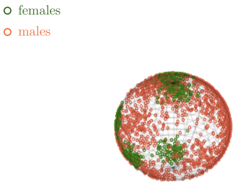
$$\mathbb{P}(X \in \mathrm{d}x) = \sum_{k=1}^{K} \pi_k \overbrace{C_d(\kappa_k) \exp\left(\kappa_k \mu_k^T x\right)}^{\text{hyperspherical gaussian}}$$

K identities
μ_k : centroid of the k-th identity
$$\kappa_k = \begin{cases} \kappa_F & \text{if female,} \\ \kappa_M & \text{if male.} \end{cases}$$

$\rightsquigarrow$ We set a mixture of von Mises-Fisher distributions, as a statistical model on the hypersphere $\mathbb{S}^{d-1}$.

The parameter κ is the inverse of the variance of a gaussian constrained to live on $\mathbb{S}^{d-1}$.

Geometric Embedding View on Fairness



$$\mathbb{P}(X \in \mathrm{d}x) = \sum_{k=1}^{K} \pi_k \overbrace{C_d(\kappa_k) \exp\left(\kappa_k \mu_k^T x\right)}^{\text{hyperspherical gaussian}}$$

K identities
μ_k : centroid of the k-th identity

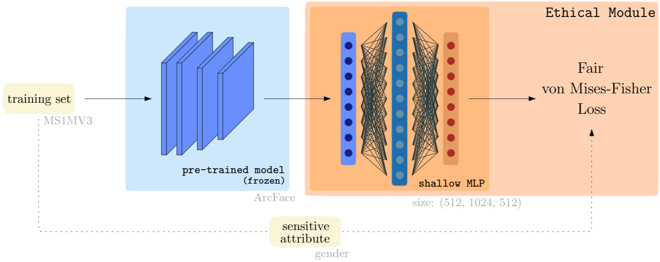
$$\kappa_k = \begin{cases} \kappa_F & \text{if female,} \\ \kappa_M & \text{if male.} \end{cases}$$

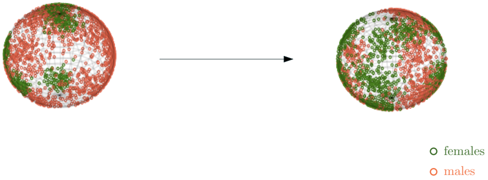
With hyperparameters κ_F and κ_M, the negative log-likelihood of the statistical model is the *Fair von Mises-Fisher loss*:

$$\mathcal{L}_{\text{FvMF}}(\Theta, \{\boldsymbol{\mu}_k\}) = -\frac{1}{N} \sum_{i=1}^{N} \log \left[\frac{C_d(\kappa_{y_i})\ e^{\kappa_{y_i}\ \boldsymbol{\mu}_{y_i}^{\mathsf{T}} \boldsymbol{z}_i}}{\sum_{k=1}^{K} C_d(\kappa_k)\ e^{\kappa_k\ \boldsymbol{\mu}_k^{\mathsf{T}} \boldsymbol{z}_i}} \right],$$

where $\boldsymbol{z}_i = f_{\Theta}(\boldsymbol{x}_i)$ is the embedding of the image $\boldsymbol{x}_i$.

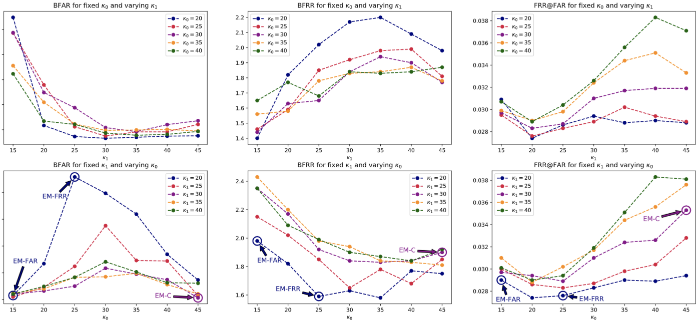
The Ethical Module





Results

BFAR and BFRR trends are correlated with κ_H and κ_F.



New SOTA for correcting the gender bias of pre-trained models

(3 methods: EM-FAR, EM-FRR, EM-C).

FAR level:	10^{-4}			10^{-3}		
Model	**FRR@FAR (%)**	**BFRR**	**BFAR**	**FRR@FAR (%)**	**BFRR**	**BFAR**
ArcFace	**0.078**	10.27	4.72	0.059	4.17	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	0.100	5.89	33.65	**0.058**	**4.11**	5.24
ArcFace + EM-C	0.164	9.18	2.44	0.081	5.15	1.20

Advantages

- Can be applied to any pre-trained model,
- Very fast training,
- Takes advantage of the performance of SOTA pre-trained networks,
- Interpretability: minimizing the Fair von Mises-Fisher loss is equivalent to maximizing the true likelihood of a Gaussian mixture model,
- The sensitive attribute (here, the gender) is only used during the training phase of the model, not afterwards.

Thanks for your attention !

For more information, please reach out to:

jean-remy.conti@telecom-paris.fr

or check out our paper