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

Jean-Rémy Conti
Télécom Paris, IDEMIA

October 12, 2022

Introduction

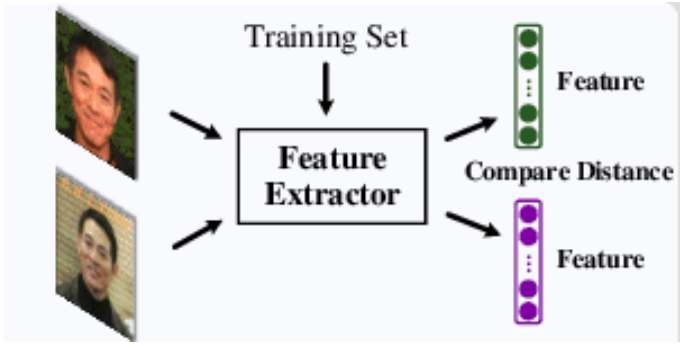
- Many applications of Face Recognition: access control, identity verification (smartphones, suspects), social media ...
- Bias with respect to race, gender, age, ...
- Different causes of bias, popular subject in machine learning

Program

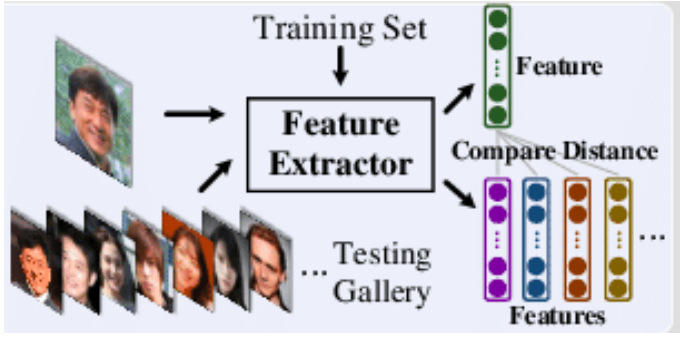
- **How to recognize a face ?**
- **How to measure fairness ?**
- **How to mitigate gender bias ?**

Face Recognition : short introduction

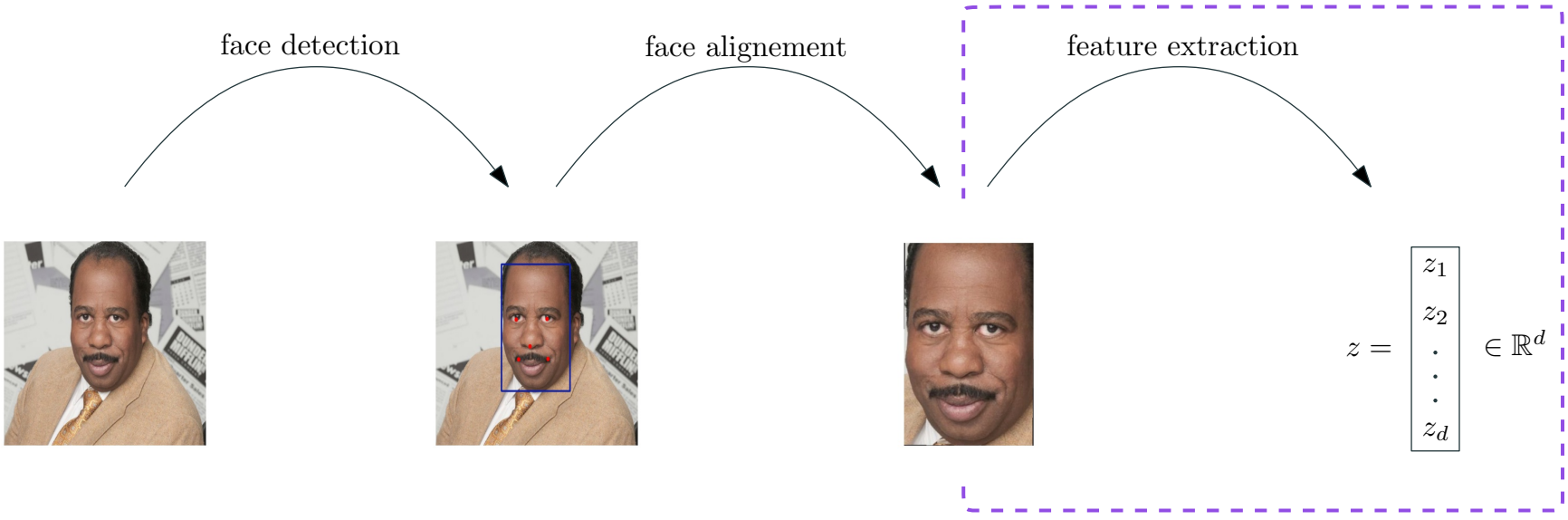
- Face verification:



- Face identification:



The steps in Face Recognition

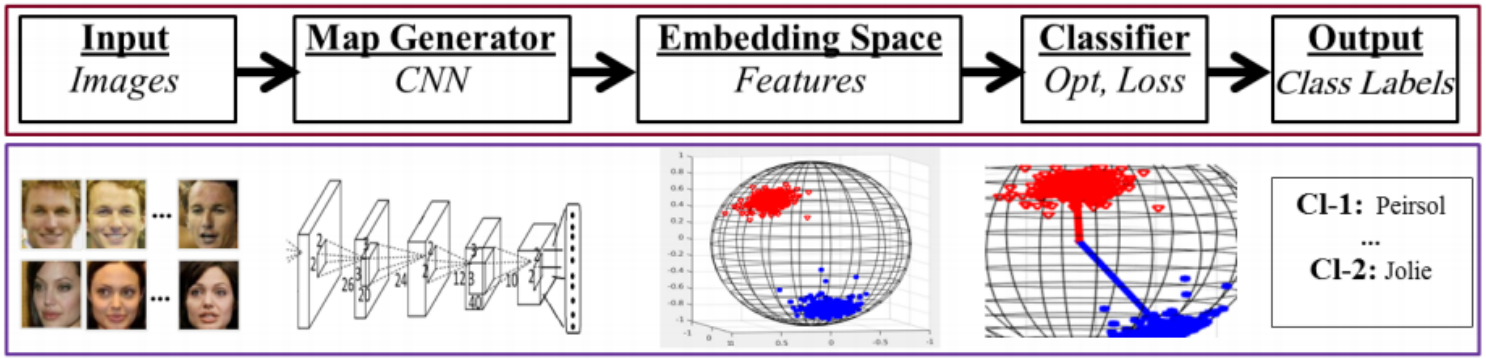


Depuis 2014 : réseau neuronal convolutif

Goal : Make the latent representations from a same identity as close as possible in the latent space.

Ingredients : training set, architecture of neural network (feature extractor), loss function.

Face Recognition training



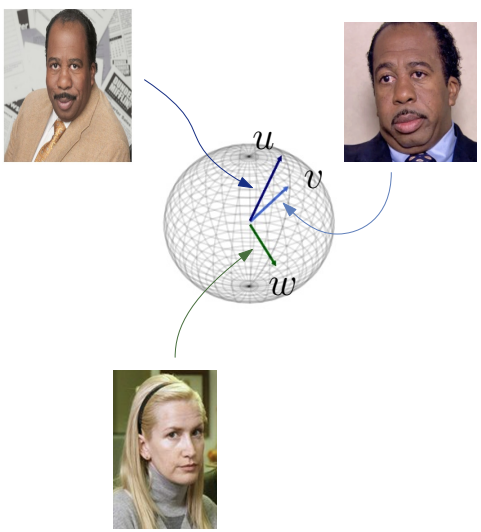
Workflow of Deep Face Recognition training.

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{\kappa \, \boldsymbol{\mu}_{y_i}{}^{\mathsf{T}} \boldsymbol{x}_i}}{\sum_{k=1}^{C} e^{\kappa \, \boldsymbol{\mu}_k{}^{\mathsf{T}} \boldsymbol{x}_i}} \qquad \|\boldsymbol{x}_i\|_2 = \|\boldsymbol{\mu}_k\|_2 = 1$$

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 $\langle u, v \rangle = \frac{\boldsymbol{u}^\intercal \boldsymbol{v}}{\|\boldsymbol{u}\|_2 \ \|\boldsymbol{v}\|_2}$.



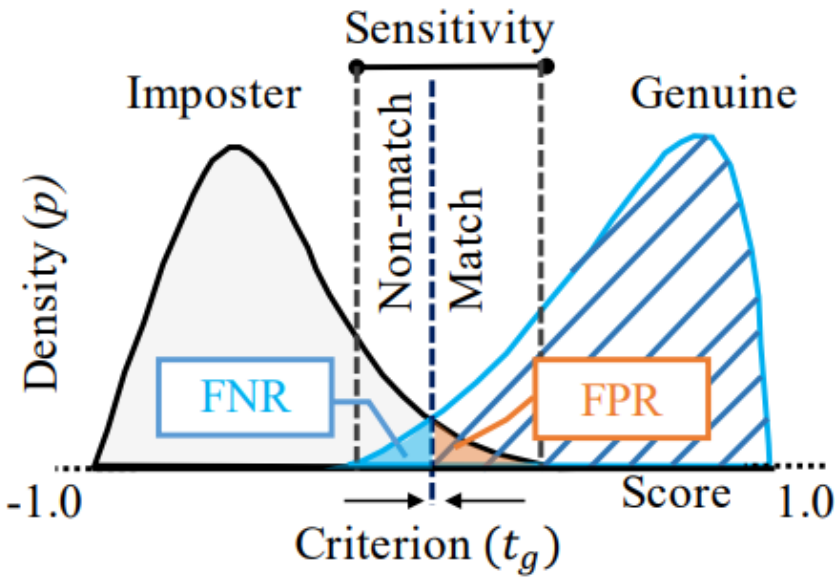
Decision rule : $t \in [-1, 1]$, fixed threshold.

- $\langle u, v \rangle \geq t \Rightarrow$ same identity (genuine),
- $\langle u, w \rangle < t \Rightarrow$ distinct identities (impostor).

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

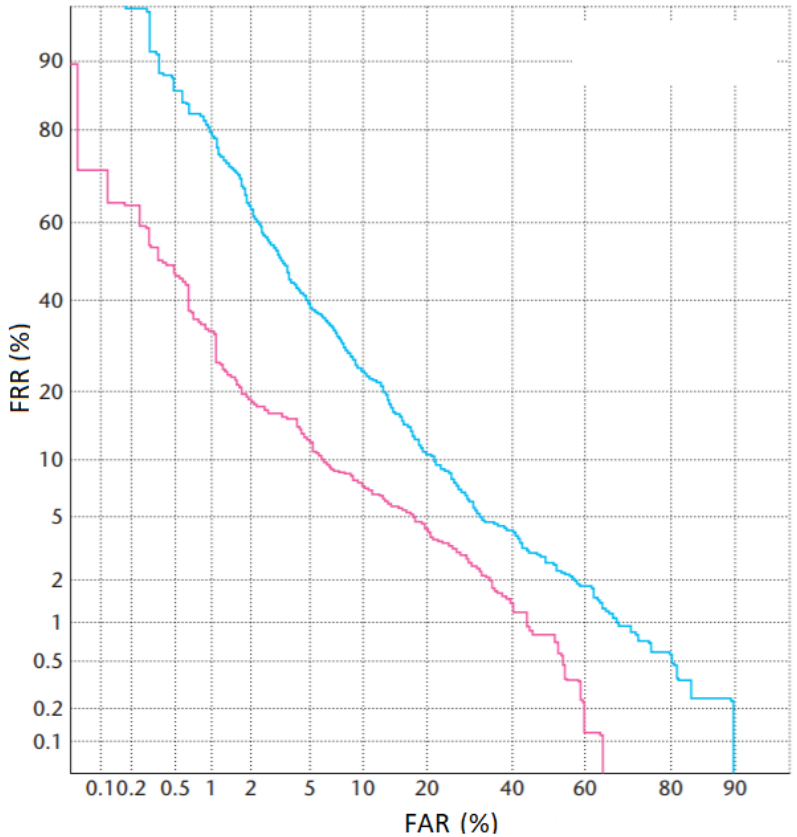
In practice :

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:

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

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

DET/ROC curve



Two typical ROC curves.

Demographic differentials

How to Measure Fairness ?

Context

Do Face Recognition algorithms have uniform performance among the population ?

- $\mathcal{G}$: 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.

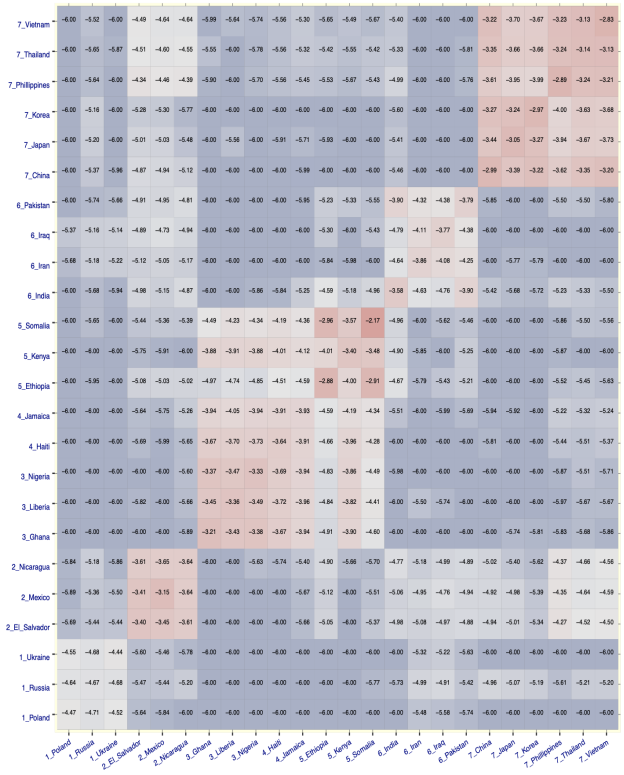
NIST reports

The *National Institute of Standards and Technology* regularly evaluates face recognition algorithms. On their performance ...

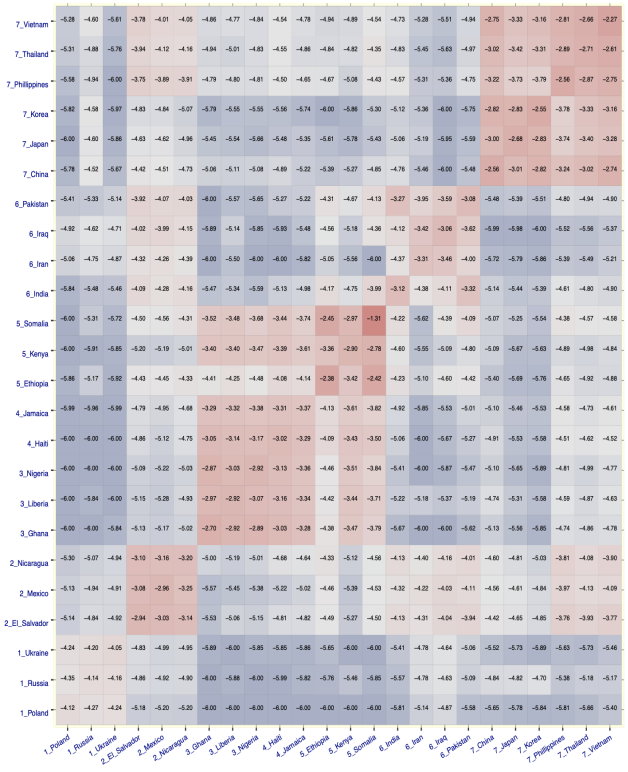
Algorithm	Constrained, Cooperative					
	FMR	= 0.000001	= 0.00001	= 0.00001	= 0.000001	= 0.000001
	Submission Date	VISA Photos	MUGSHOT Photos	MUGSHOT Photos 12+YRS	VISABORDER Photos	BORDER Photos
sensetime-005	2021-05-24	0.0029(17)	0.0022(1)	0.0021(1)	0.0023(1)	0.0044(1)
visionlabs-011	2021-10-13	0.0022(7)	0.0024(9)	0.0026(7)	0.0028(2)	0.0053(2)
ntechlab-011	2021-09-13	0.0019(4)	0.0024(8)	0.0028(18)	0.0029(3)	0.0055(3)
clearviewai-000	2021-09-22	0.0019(5)	0.0024(5)	0.0028(11)	0.0030(4)	0.0058(5)
mendaxiatech-000	2021-09-15	0.0036(25)	0.0029(26)	0.0036(32)	0.0031(5)	0.0057(4)
ntechlab-010	2021-04-30	0.0017(2)	0.0024(10)	0.0029(20)	0.0031(6)	0.0058(6)
cubox-002	2021-08-24	0.0041(35)	0.0025(11)	0.0025(6)	0.0033(7)	0.0064(9)
visionlabs-010	2021-01-25	0.0024(9)	0.0026(16)	0.0030(22)	0.0033(8)	0.0061(7)
toshiba-004	2021-09-27	0.0042(40)	0.0025(12)	0.0027(9)	0.0034(9)	0.0063(8)
idemia-008	2021-07-07	0.0032(19)	0.0023(3)	0.0028(10)	0.0034(10)	0.0067(11)
kakao-006	2021-10-13	0.0029(16)	0.0024(4)	0.0028(16)	0.0035(11)	0.0065(10)
insightface-001	2021-09-27	0.0014(1)	0.0027(19)	0.0024(3)	0.0035(12)	0.0070(13)
paravision-008	2021-07-01	0.0025(11)	0.0024(6)	0.0025(5)	0.0036(13)	0.0070(14)
insightface-000	2021-03-17	0.0027(15)	0.0029(25)	0.0030(23)	0.0038(14)	0.0077(20)

NIST reports

... And on their performance differentials with respect to some subgroups of the population !



FAR for men.



FAR for women.

NIST reports

... And on their performance differentials with respect to some subgroups of the population !

FAR for ethnicity+gender subgroups.
F: female, M: male, B: black, W: white.

$\rightsquigarrow$ Some algorithms make 10 times more errors on black women than on white men.

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:

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

[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.

Bias Mitigation in Face Recognition

Survey of existing methods

Pre-training : reweighting / augmentation

- **Balanced Datasets Are Not Enough**, Wang and al. 2019.
- **How Does Gender Balance In Training Data Affect Face Recognition Accuracy?**, Albiero and al. 2020.

⇝ Not yet adapted to Face Recognition.

Adversarial methods during training

- **PASS: Protected attribute suppression system for mitigating bias in face recognition**, Dhar and al. 2021.
- **How Does Gender Balance In Training Data Affect Face Recognition Accuracy?**, Albiero and al. 2020.

⇝ Costly in computing time and unstable.

Post-training methods : modification of matching scores

- **Bias mitigation of face recognition models through calibration.**, Salvador and al. 2021.

⇝ Does not solve the problem at the root level.

Our proposition

Core idea : add a shallow neural network on the last layer of a pre-trained model, in order to correct its gender bias.

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

---

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

Accepted at ICML 2022 conference.

Geometric Embedding View on Fairness

Women are disadvantaged compared to men in terms of both FAR and FRR.

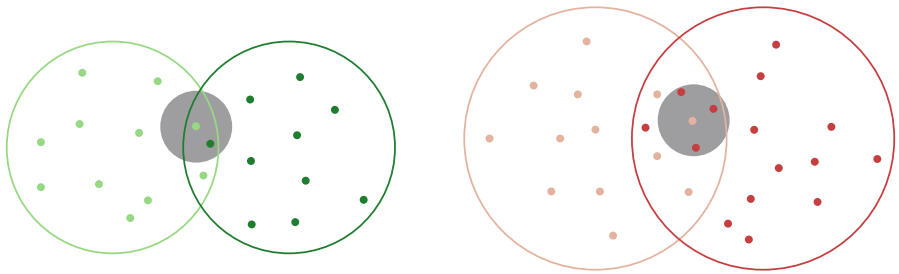


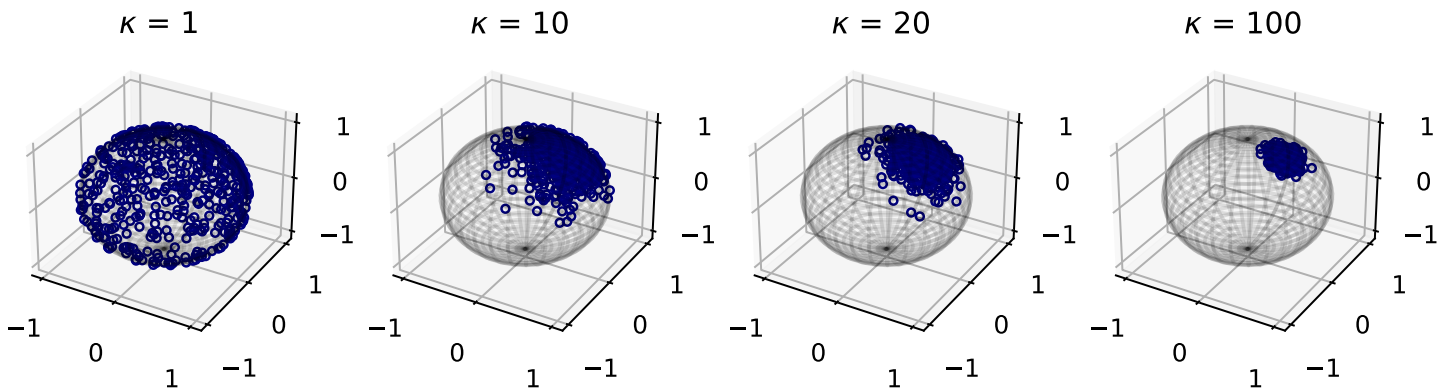
Illustration of the geometric nature of bias. Each point is the embedding of an image. In green: two male identities. In red: two female identities. The overlapping region between two identities is higher for females than for males.

$\rightsquigarrow$ We choose to change the spread of each identity, according to their gender.

vMF distribution

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

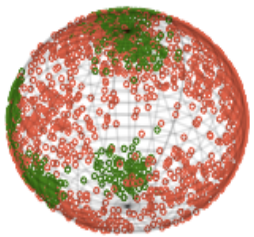
$$V_d(\boldsymbol{z}; \boldsymbol{\mu}, \kappa) := C_d(\kappa) e^{\kappa \boldsymbol{\mu}^\mathsf{T} \boldsymbol{z}},$$



500 samples from the vMF distribution in dimension 3.

Statistical Model on the Hypersphere

females
males



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

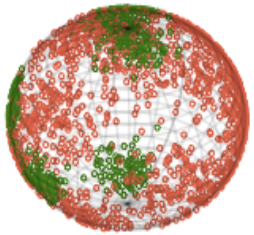
K identities
μ_k : centroid of the k-th identity
$\kappa_k = \begin{cases} \kappa_F & \text{if female,} \\ \kappa_M & \text{if male.} \end{cases}$

$\rightsquigarrow$ 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 $\mathbb{S}^{d-1}$.

Statistical Model on the Hypersphere

females
males



hyperspherical gaussian

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

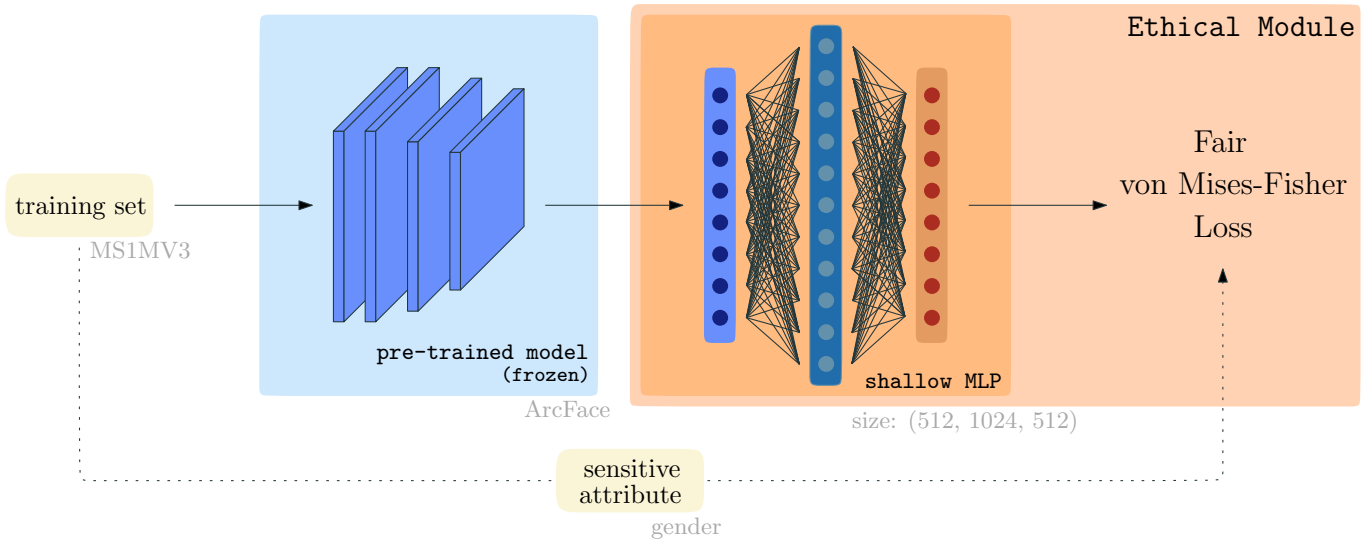
K identities
μ_k : centroid of the k-th identity
$\kappa_k = \begin{cases} \kappa_F & \text{if female,} \\ \kappa_M & \text{if male.} \end{cases}$

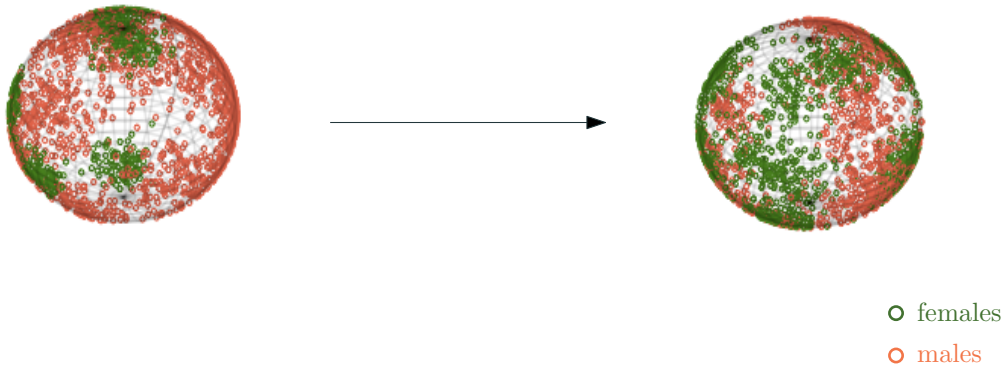
With hyperparameters κ_F and κ_M, the negative log-likelihood of the statistical model is the *Fair von Mises-Fisher loss*:

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

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

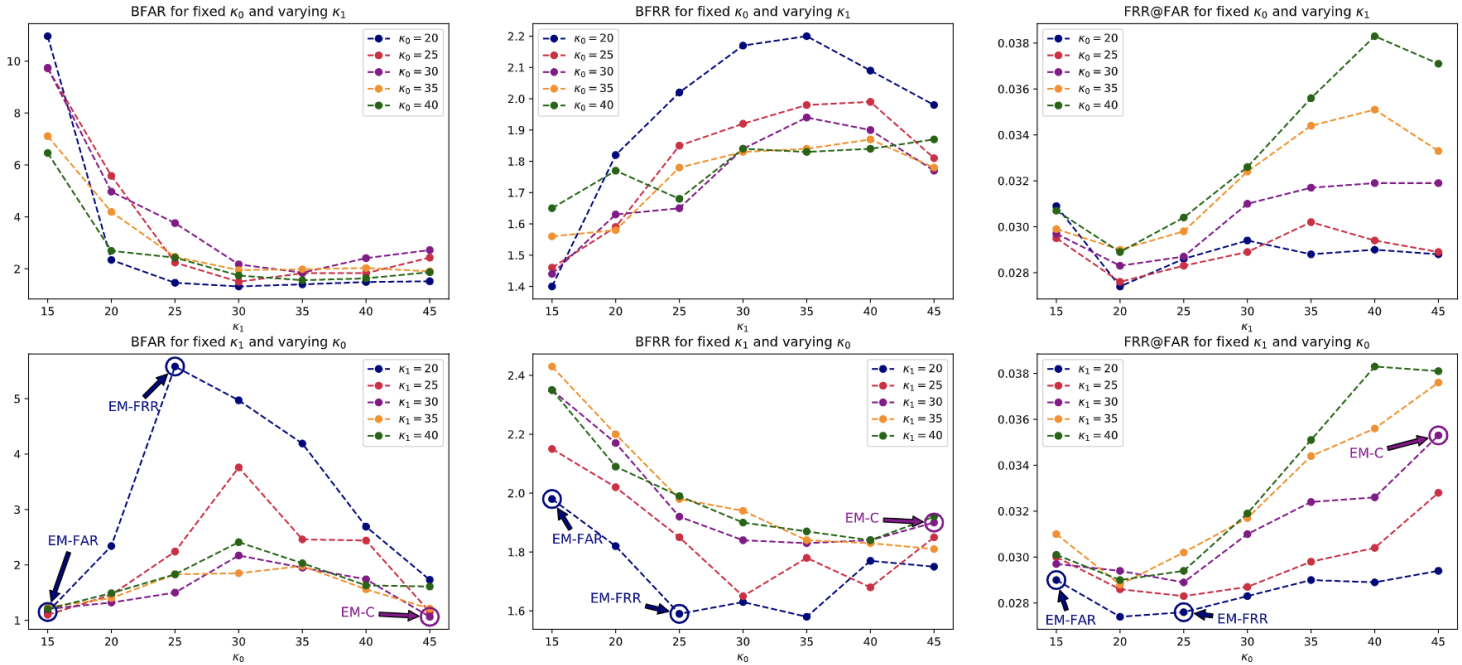
The Ethical Module





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