Learning an Ethical Module for Bias Mitigation of pre-trained Face Recognition Models

Journée de la chaire DSAIDIS

Jean-Rémy Conti, *IDEMIA, Télécom Paris*
Nathan Noiry, *Télécom Paris*

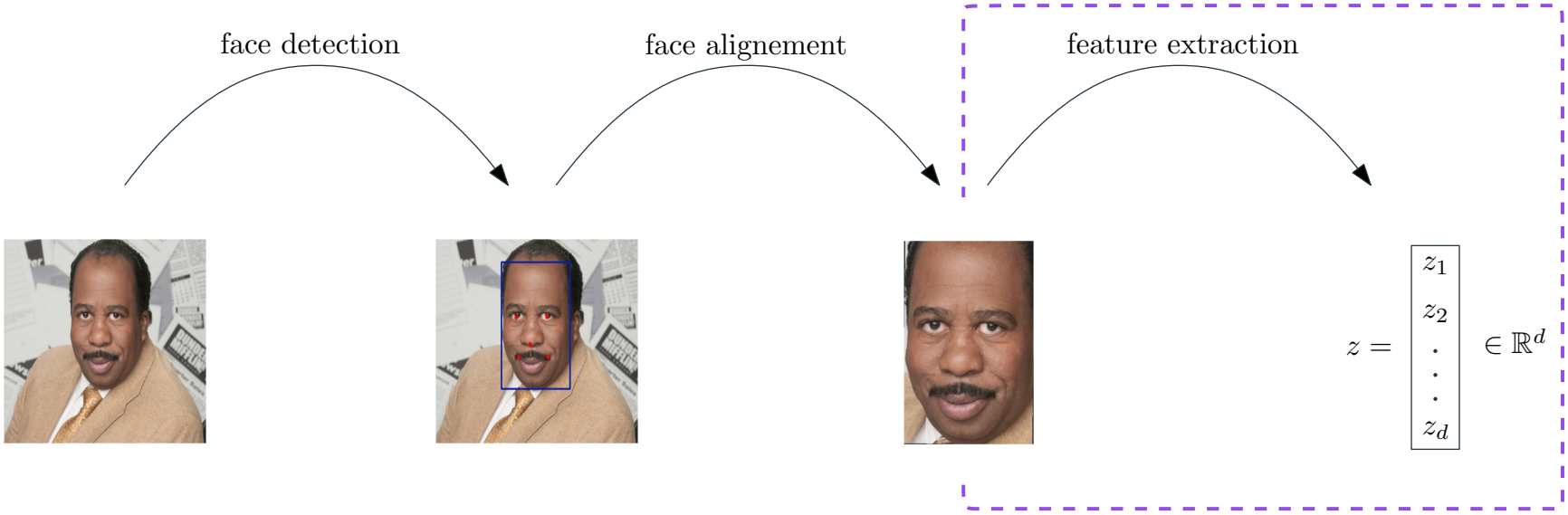
Télécom Paris
15 juin 2022

Au programme

- **Comment reconnaître un visage ?**
- **Comment mesurer l'équité ?**
- **Comment corriger un biais de genre ?**

Reconnaissance faciale : introduction rapide

Les étapes de la reconnaissance faciale



Depuis 2014 : réseau neuronal convolutif

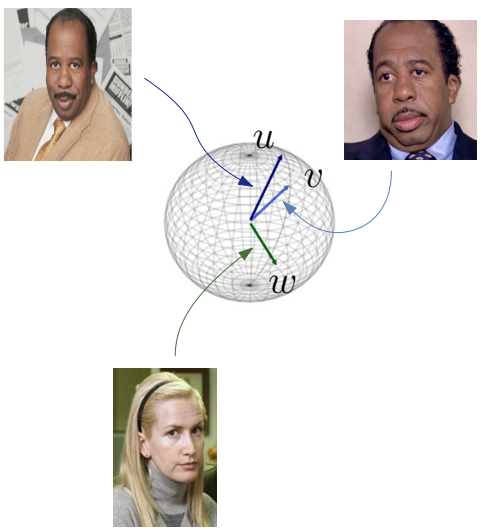
Objectif : Faire en sorte que les différentes représentations d'une même identité soient proches dans l'espace latent.

Ingrédients : fonction de perte, base de données, architecture.

Les étapes de la reconnaissance faciale

Dernière étape : comparer deux visages *(face matching)*.

La majorité des systèmes utilisent des représentations normalisées sur l'hypersphère $\mathbb{S}^{d-1}$ et mesurent la similarité avec le produit scalaire usuel.



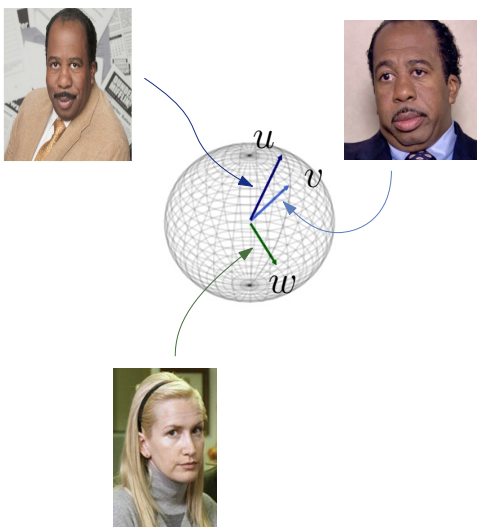
Règle de décision : $t \in [-1, 1]$ seuil fixé.

- $\langle u, v \rangle \geq t \Rightarrow$ "même identité",
- $\langle u, w \rangle < t \Rightarrow$ "identités distinctes".

Les étapes de la reconnaissance faciale

Dernière étape : comparer deux visages *(face matching)*.

La majorité des systèmes utilisent des représentations normalisées sur l'hypersphère $\mathbb{S}^{d-1}$ et mesurent la similarité avec le produit scalaire usuel.



Règle de décision : $t \in [-1, 1]$ seuil fixé.

- $\langle u, v \rangle \geq t \Rightarrow$ "même identité",
- $\langle u, w \rangle < t \Rightarrow$ "identités distinctes".

Deux applications :

- Vérification (1:1) : est-ce que deux images correspondent à la même personne ?
  $\rightsquigarrow$ contrôle aux frontières.
- Reconnaissance (1:N) : est-ce qu'une image donnée correspond à l'identité d'une base existante ?
  $\rightsquigarrow$ recherche d'individu.

Métrique d'évaluation

Deux types d'erreur peuvent se produire :

- Faux positifs : prédire "même identité" pour une paire d'images de deux individus distincts. $\rightsquigarrow$ False Acceptance Rate: FAR.
- Faux négatifs : prédire "identités distinctes" pour une paire d'images d'un même individu. $\rightsquigarrow$ False Rejection Rate: FRR.

Métrique d'évaluation

Deux types d'erreur peuvent se produire :

- Faux positifs : prédire "même identité" pour une paire d'images de deux individus distincts. $\rightsquigarrow$ False Acceptance Rate: FAR.
- Faux négatifs : prédire "identités distinctes" pour une paire d'images d'un même individu. $\rightsquigarrow$ False Rejection Rate: FRR.

En pratique :

1. On fixe un point de fonctionnement $t \in [-1, 1]$ afin d'obtenir un taux de faux positifs α jugé acceptable.
2. On calcule le taux de faux négatifs à ce seuil.

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

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

Rapports du NIST

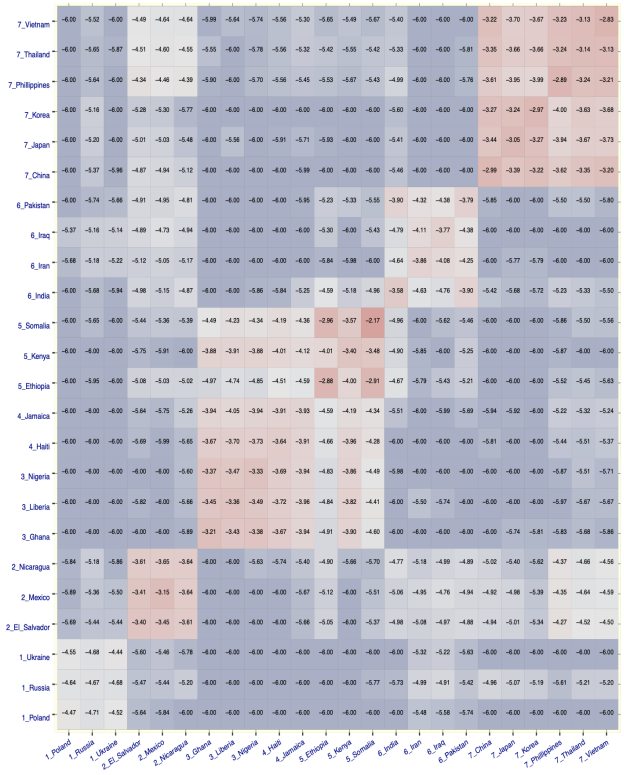
Le *National Institute of Standards and Technology* évalue régulièrement les algorithmes de reconnaissance faciale. Sur la performance...

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)

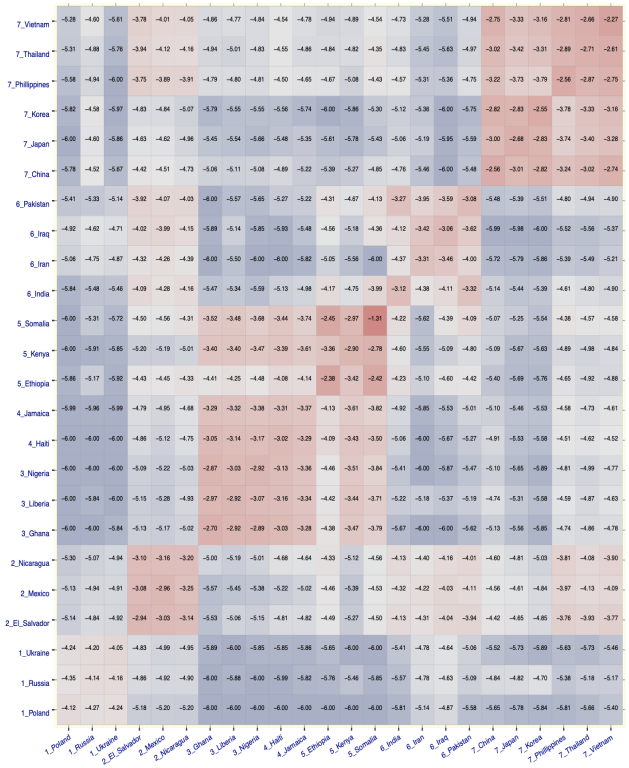
Différentiel démographique

Rapports du NIST

... Et sur les disparités de performance en fonction des sous-groupes de la population !



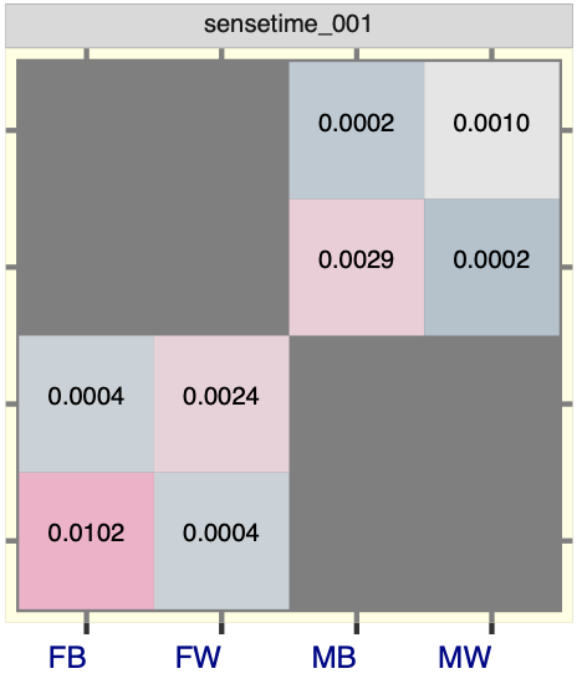
FAR pour les hommes.



FAR pour les femmes.

Rapports du NIST

... Et sur les disparités de performance en fonction des sous-groupes de la population !



FAR pour des sous-groupes ethnie+genre.
F: female, M: male, B: black, W: white.

$\rightsquigarrow$ Certains algorithmes commettent 10 fois plus d'erreurs sur les femmes noires que sur les hommes blancs.

Comment mesurer l'équité

Contexte

- $\mathcal{G}$: ensemble de sous-groupes de la population.
  Exemples : hommes / femmes, jeunes / âgés ...
- Pour tout $g \in \mathcal{G}$, FAR_g et FRR_g False Acceptance and False Rejection Rates au sein du sous-groupe g.

Comment mesurer l'équité

Contexte

- $\mathcal{G}$: ensemble de sous-groupes de la population.
  Exemples : hommes / femmes, jeunes / âgés ...
- Pour tout $g \in \mathcal{G}$, FAR_g et FRR_g False Acceptance and False Rejection Rates au sein du sous-groupe g.

Plusieurs propositions de métriques

1. Deux propositions de ratios :

$$\underbrace{\frac{\max_g \mathrm{FAR}_g(t)}{\min_g \mathrm{FAR}_g(t)}}_{\text{pire écart}} \quad \text{v.s.} \quad \underbrace{\frac{\max_{g\in\mathcal{G}} \mathrm{FAR}_g(t)}{(1/|\mathcal{G}|)\sum_g \mathrm{FAR}_g(t)}}_{\text{écart à la moyenne}}.$$

Comment mesurer l'équité

Contexte

- $\mathcal{G}$: ensemble de sous-groupes de la population.
  Exemples : hommes / femmes, jeunes / âgés ...
- Pour tout $g \in \mathcal{G}$, FAR_g et FRR_g False Acceptance and False Rejection Rates au sein du sous-groupe g.

Plusieurs propositions de métriques

1. Deux propositions de ratios :

$$\underbrace{\frac{\max_g \mathrm{FAR}_g(t)}{\min_g \mathrm{FAR}_g(t)}}_{\text{pire écart}} \quad \text{v.s.} \quad \underbrace{\frac{\max_{g\in\mathcal{G}} \mathrm{FAR}_g(t)}{(1/|\mathcal{G}|)\sum_g \mathrm{FAR}_g(t)}}_{\text{écart à la moyenne}}.$$

2. Deux propositions de seuil :

$$\mathrm{FAR}(t) = \alpha \quad \text{v.s.} \quad \max_{g\in\mathcal{G}} \mathrm{FAR}_g(t) = \alpha.$$

Correction des biais de genre

Survol de quelques méthodes existantes

Pré-entraînement : re-pondération / 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.

⇝ Pas encore adapté pour la reconnaissance faciale.

Méthodes adverses pendant l'entraînement

- **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.

⇝ Prohibitif en temps de calcul et instable.

Méthodes de post-traitement : modification des scores de similarité

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

⇝ Ne résout pas le problème à la source.

Notre proposition

Notre proposition : greffer un réseau de neurone superficiel sur la dernière couche d'un modèle pré-entraîné afin de corriger les biais de représentativité de l'espace latent.

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

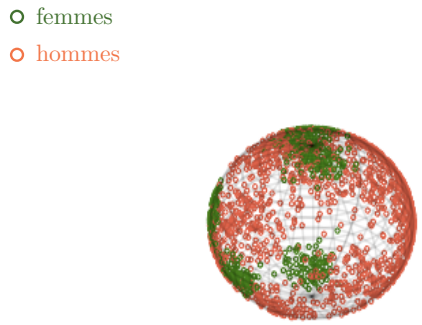
---

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

Accepté pour publication à ICML 2022.

Origine géométrique des biais et modèle statistique

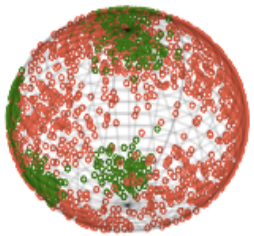
Observation : Les représentations latentes des femmes occupent une portion plus petite de l'hypersphère que les représentations latentes des hommes.



Origine géométrique des biais et modèle statistique

Observation : Les représentations latentes des femmes occupent une portion plus petite de l'hypersphère que les représentations latentes des hommes.

femmes
hommes



$$\mathbb{P}(X \in \mathrm{d}x) = \sum_{k=1}^{K} \pi_k \overbrace{C_d(\kappa_k) \exp\left(\kappa_k \mu_k^T x\right)}^{\text{gaussienne sur l'hypersphère}}$$

K identités
μ_k : centroid de l'identité k

$$\kappa_k = \begin{cases} \kappa_F & \text{si femme,} \\ \kappa_H & \text{si homme.} \end{cases}$$

$\rightsquigarrow$ Mélange de von-Mises Fisher. Le paramètre κ est appelé coefficient de concentration et peut être interprété comme l'inverse d'une variance.

Maximum de vraissemblance et loss de vMF

Si l'on fixe les coefficients de concentrations κ_H et κ_F, la log-vraissemblance ℓ du modèle s'écrit

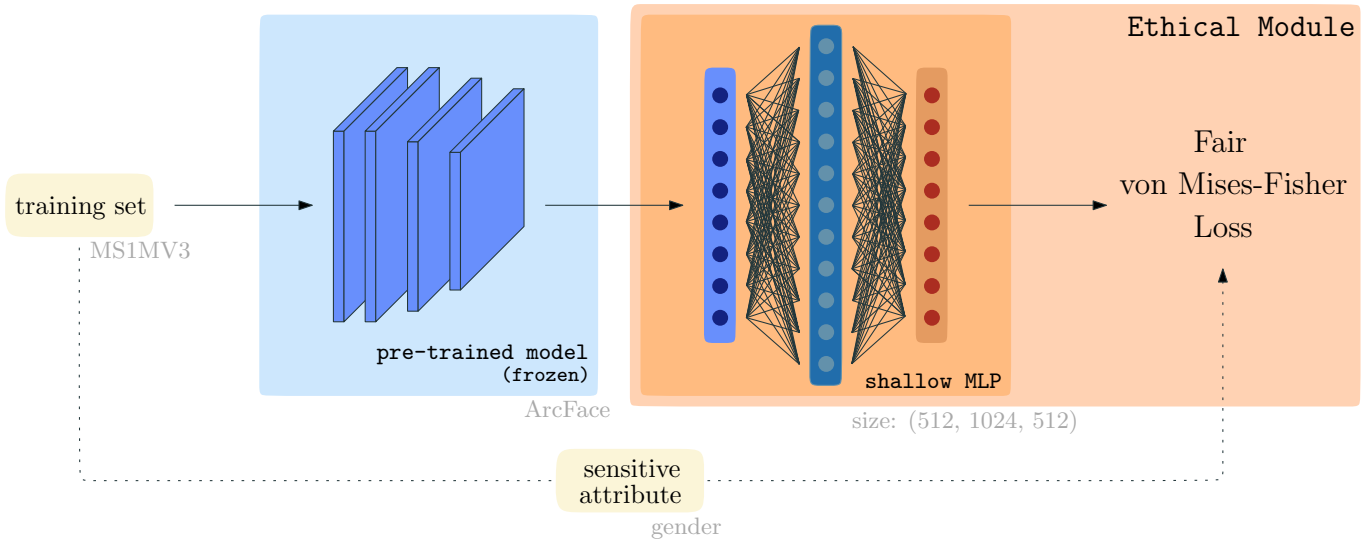
$$\ell(\mu, \theta) = \frac{1}{N} \sum_{i=1}^{N} \log \left[\frac{C_d(\kappa_{y_i})\ e^{\kappa_{y_i}\ \mu_{y_i}^{\mathsf{T}} z_i}}{\sum_{k=1}^{K} C_d(\kappa_k)\ e^{\kappa_k\ \mu_k^{\mathsf{T}} z_i}} \right],$$

où $z_i = f_\theta(x_i)$ est la représentation latente d'une image x_i.

On introduit ainsi une *Fair von Mises-Fisher loss* :

$$\mathcal{L}_{\text{FvMF}} := -\ell(\mu, \theta).$$

Description du Module Ethique



femmes

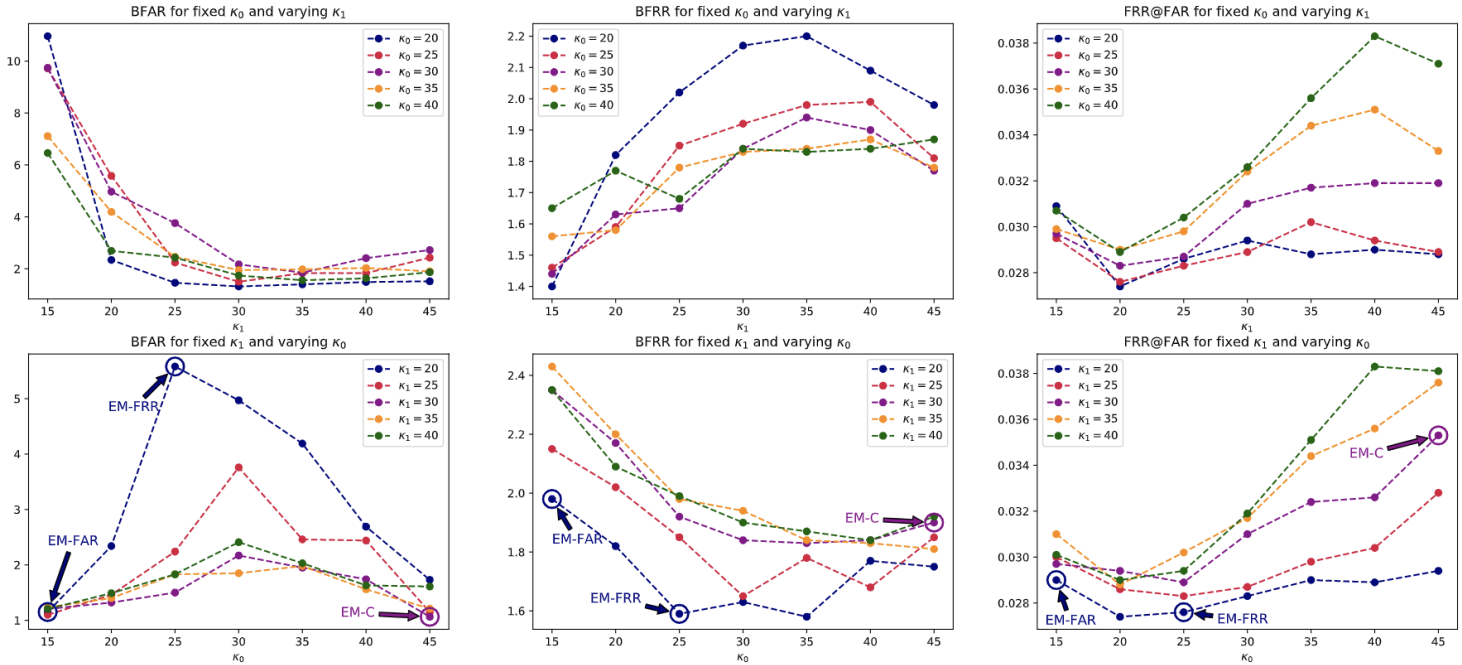
hommes

Avantages de la méthode

- Application à tout réseau pré-entraîné,
- Entraînement rapide,
- Tire partie de la performance des réseaux pré-entraînés,
- Interprétabilité : minimiser la loss de von Mises-Fisher revient à maximiser la vraissemblance d'un modèle de mélange gaussien,
- L'attribut sensible est uniquement utilisé lors de la phase d'entraînement du modèle, pas au cours de son fonctionnement en production.

Résultats prometteurs

BFAR and BFRR trends are correlated with κ_H and κ_F.



Amélioration de l'état de l'art pour certains points choisis.

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

Travail à venir

- Adaptation de la méthode dans d'autres contextes que la reconnaissance faciale
- Adaptation de la méthode pour d'autres problèmes d'équités en reconnaissance faciale (race, âge)
- Mise au point d'un algorithme d'exploration efficace de l'espace des hyperparamètres.

...

D'autres suggestions ?

Merci de votre attention !