Rethinking Bias Mitigation: Fairer Architectures Make for Fairer Face Recognition (NeurIPS 2023)

Samuel Dooley, Rhea Sanjay Sukthanker, John P. Dickerson, Colin White, Frank Hutter, Micah Goldblum

Yuha Park

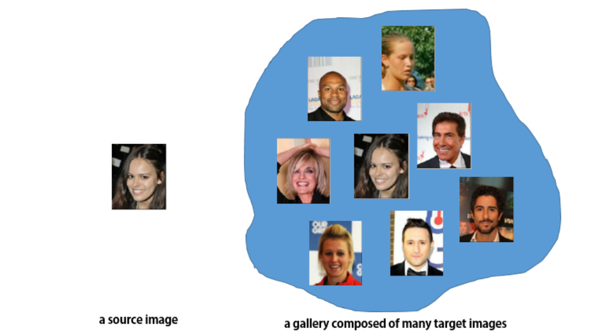
January 8, 2024

Seoul National University

Contents

Introduction

Introduction



- **Face identification tasks** ask whether a given person in a source image appears within a gallery composed of many target images (one-to-many comparision).
- Face recognition models exhibit bias, such as gender and race.
- Conventional wisdom dictates that model biases arise from biased training data.
- A fundamental question: *Does model bias arise from the architecture and hyperparameters?*

  $\Longrightarrow$ **Neural Architecture Search (NAS)** $\times$ **Hyperparameter Optimization(HPO)**

Neural Architecture Search (NAS) & Hyperparameter Optimization (HPO)

- **NAS** aims at automating the design of network architectures.
- **HPO** refers to the automated search for optimal hyperparameters.

  (learning rate, batch size, dropout, loss function, optimizer, and architectural choices, etc.)
- **Limitaions of existing studies in face recognition systems**
  - The training hyperparameters for the architectures are *fixed* in NAS techniques.
  - None of the methods can be applied for a *joint* architecture and hyperparameter search.
  - None of them have been used to *optimize fairness*.

Are Architectures and Hyperparameters Important for Fairness?

Evaluation metric

- **Error** (representation error): for a given image, whether the closest image in feature space is *not* of the same person based on ℓ_2 distance.
- **Rank**: how many images of a different identity are closer to the image in feature space.

$$\implies \mathit{Rank}(\mathit{image}) = 0 \text{ iff } \mathit{Error}(\mathit{image}) = 0; \mathit{Rank}(\mathit{image}) > 0 \text{ iff } \mathit{Error}(\mathit{image}) = 1.$$

- **Rank disparity**:

$$\left| \frac{1}{|\mathcal{D}_{\text{male}}|} \sum_{x \in \mathcal{D}_{\text{male}}} \mathit{Rank}(x) - \frac{1}{|\mathcal{D}_{\text{female}}|} \sum_{x \in \mathcal{D}_{\text{female}}} \mathit{Rank}(x) \right|.$$

Result

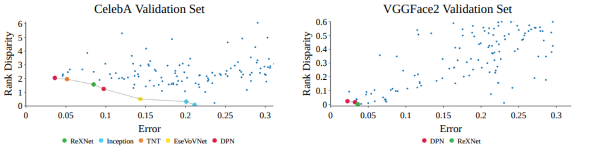
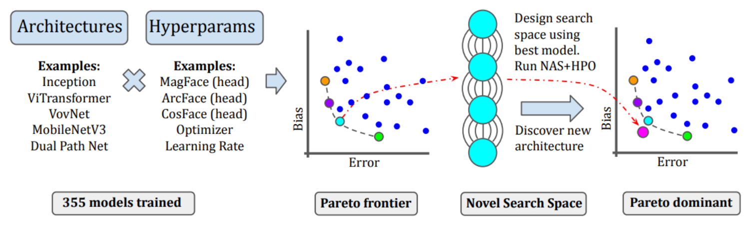


Figure 2: (Left) CelebA (Right) VGGFace2. Error-Rank Disparity Pareto front of the architectures with lowest error (< 0.3). Models in the lower left corner are better. The Pareto front is denoted with a dashed line. Other points are architecture and hyperparameter combinations which are not Pareto-optimal.

- Optimizing for error does not always optimize for fairness.
- Different architectures have different fairness properties.
- DPN architecture has the lowest error and is Pareto-optimal on both datasets.
- There are differences between the two datasets at the most extreme low errors.
  - For VGGFace2, there are 10 models with $Error < 0.05$; CelebA has 3 such models.
  - Models with low error also have low rank disparity on VGGFace2 but *not* for CelebA.
  - The Pareto-optimal models differ across datasets.
  - Different architectures exhibit different Pareto-optimal hyperparameters.

Neural Architecture Search for Bias Mitigation

Overview of methodology



Search Space Design

Index	Operation	Definition
0	BnConv1x1	Batch Normalization $\rightarrow$ Convolution with 1x1 kernel
1	Conv $1\mathrm{x}1\mathrm{Bn}$	Convolution with 1x1 kernel $\rightarrow$ Batch Normalization
2	Conv1x1	Convolution with 1x1 kernel
3	BnConv3x3	Batch Normalization $\rightarrow$ Convolution with 3×3 kernel
4	Conv $3 \times 3Bn$	Convolution with 3×3 kernel $\rightarrow$ Batch Normalization
5	Conv 3×3	Convolution with 3×3 kernel
6	BnConv5x5	Batch Normalization $\rightarrow$ Convolution with 5×5 kernel
7	Conv $5 \times 5Bn$	Convolution with 5×5 kernel $\rightarrow$ Batch Normalization
8	Conv5x5	Convolution with 5×5 kernel

Table 1: Operation choices (Architecture).

Hyperparameter	Choices
Architecture Head/Loss	MagFace, ArcFace, CosFace
Optimizer Type	Adam, AdamW, SGD
Learning rate (conditional)	Adam/AdamW $\rightarrow [1e-4, 1e-2]$, SGD $\rightarrow [0.09, 0.8]$

Table 2: Searchable hyperparameter choices.

Results

Table 1: Comparison of bias mitigation techniques where the SMAC models were found with our NAS+HPO bias mitigation technique and the other three techniques are standard in facial recognition: Flipped [9], Angular [76], and SensitiveNets [110]. Items in bold are Pareto-optimal. The values show (Error;Rank Disparity). Other metrics are reported in Appendix C.6 and Table 8.

	Trained on VGGFace2					Trained on CelebA			
Model	**Baseline**	**Flipped**	**Angular**	**SensitiveNets**	**Model**	**Baseline**	**Flipped**	**Angular**	**SensitiveNets**
SMAC_301	**(3.66;0.23)**	**(4.95;0.18)**	(4.14;0.25)	(6.20;0.41)	SMAC_000	(3.25;2.18)	**(5.20;0.03)**	(3.45;2.28)	(3.45;2.18)
DPN	(3.56;0.27)	(5.87;0.32)	(6.06;0.36)	(4.76;0.34)	SMAC_010	(4.14;2.27)	(12.27; 5.46)	(4.50;2.50)	(3.99;2.12)
ReXNet	(4.09;0.27)	(5.73;0.45)	(5.47;0.26)	(4.75;0.25)	SMAC_680	**(3.22;1.96)**	(12.42;4.50)	(3.80;4.16)	(3.29;2.09)
Swin	(5.47;0.38)	(5.75;0.44)	(5.23;0.25)	(5.03;0.30)	ArcFace	(11.30;4.6)	(13.56;2.70)	(9.90;5.60)	(9.10;3.00)

Table 2: We transfer the evaluation of top performing models on VGGFace2 and CelebA onto six other common face recognition datasets: LFW [53], CFP_FF [100], CFP_FP [100], AgeDB [77], CALFW [128], CPLPW [127]. The novel architectures found with our bias mitigation strategy significantly outperform other models in terms of accuracy. Refer Table 9 for the complete results.

Architecture (trained on VGGFace2)	**LFW**	**CFP_FF**	**CFP_FP**	**AgeDB**	**CALFW**	**CPLFW**
Rexnet_200	82.60	80.91	65.51	59.18	68.23	62.15
DPN_SGD	93.0	91.81	78.96	71.87	78.27	72.97
DPN_AdamW	78.66	77.17	64.35	61.32	64.78	60.30
SMAC_301	**96.63**	**95.10**	**86.63**	**79.97**	**86.07**	**81.43**
Architecture (trained on CelebA)	**LFW**	**CFP_FF**	**CFP_FP**	**AgeDB**	**CALFW**	**CPLFW**
DPN_CosFace	87.78	90.73	69.97	65.55	75.50	62.77
DPN_MagFace	91.13	92.16	70.58	68.17	76.98	60.80
SMAC_000	**94.98**	95.60	**74.24**	80.23	84.73	64.22
SMAC_010	94.30	94.63	73.83	**80.37**	84.73	**65.48**
SMAC_680	94.16	**95.68**	72.67	79.88	**84.78**	63.96