Presented By:
Sreeraj Ramachandran

Leveling Down in Computer Vision: Pareto Inefficiencies in Fair Deep Classifiers

Authors: Dominik Zietlow, Michael Lohaus, Guha Balakrishnan, Matthäus Kleindessner, Francesco Locatello, Bernhard Schölkopf, Chris Russell

Introduction

Unfairness?

Systematic accuracy differences across protected subgroups

Quantifying Unfairness

Comparing Accuracy-related rates between groups. E.g., Difference of Equal Opportunity (DEO) compares Groupwise True Positive Rates.

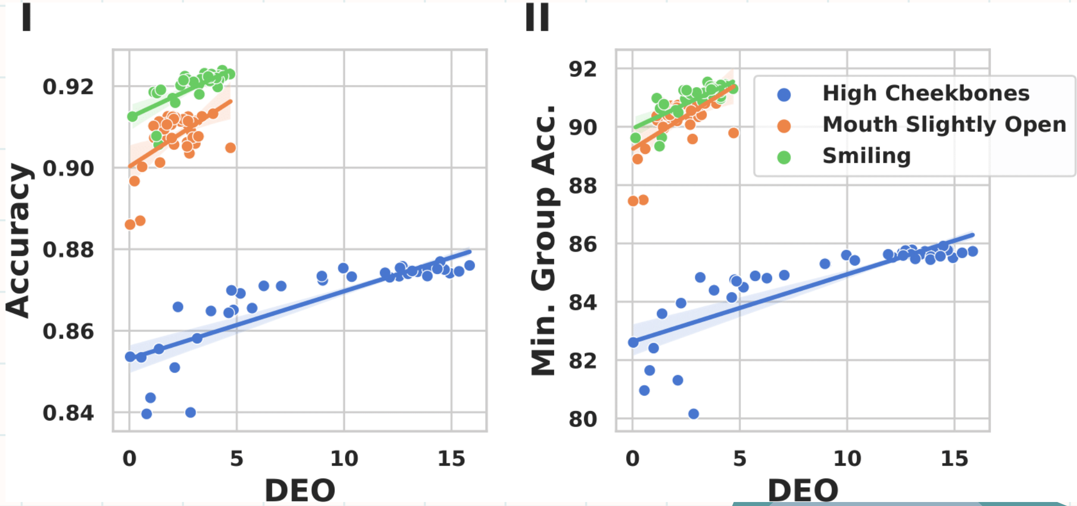
Balancing Fairness and Accuracy in Computer Vision

Explores trade-off between fairness and accuracy between better-performing and worse-performing groups in low capacity models

Introduction

Balancing Fairness and Accuracy in Computer Vision

Correlation Between Fairness and Accuracy on varying strength of fairness regularizer

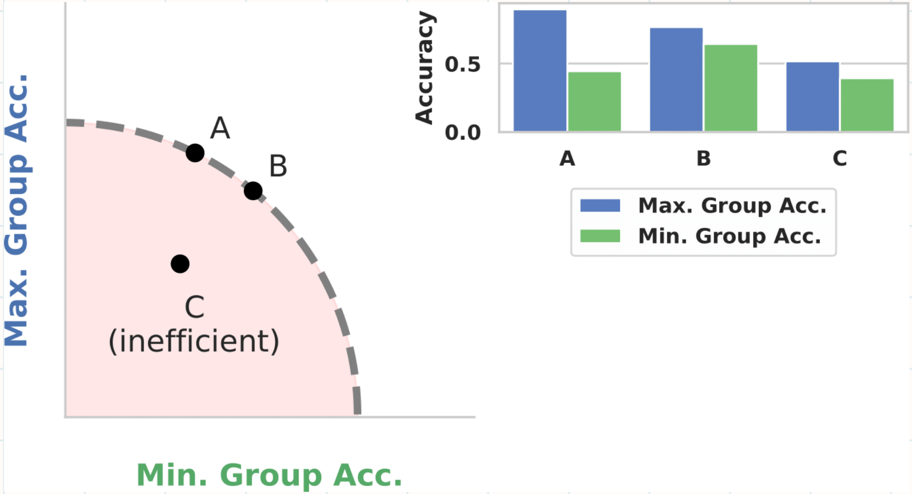


Pareto Inefficiency

High Capacity Classifiers in Computer Vision

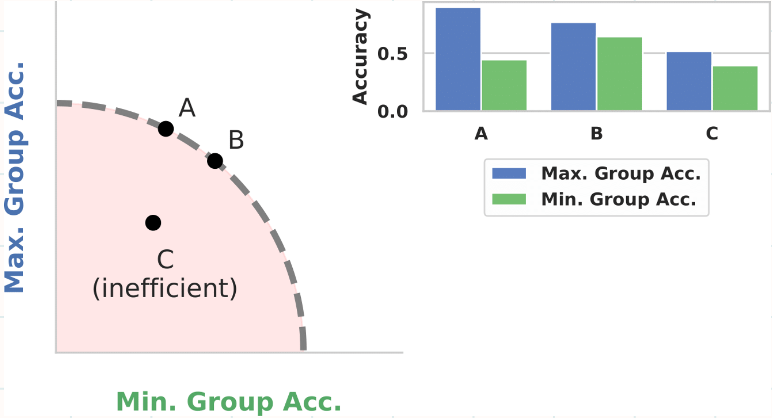
- **Degrades** Accuracy of all groups
- Increases Fairness at the cost of **worse-performing classifier** (Model C)
- Balancing fairness by **degrading** better-performing groups – **Levelling Down**

Pareto Curve



Pareto Inefficiency

Pareto Curve



Problems

- Fairness methods that decrease performance for all groups, making them **Pareto Inefficient**, should be avoided when group accuracy is a primary concern.
- High-capacity classifierss **fit training data nearly perfectly**
- **Inappropriate evaluation** of fairness methods

Notions of Fairness

Let A be set of protected attributes, $Y \in \{0,1\}$ ground-truth label and $\hat{Y} \in \{0,1\}$ the prediction

Difference in Equal Opportunity

For two groups $a, a' \in A$ violation of equal opportunity is measured by the difference in equal opportunity (DEO) defined as

$$|P(\hat{Y} = 1 | Y = 1, A = a) - P(\hat{Y} = 1 | Y = 1, A = a')|$$

Difference in Equalized Odds (DEOdds)

$$\sum_y |P(\hat{Y} = 1 | Y = y, A = a) - P(\hat{Y} = 1 | Y = y, A = a')|$$

Min-Max Fairness

Decrease the classification error for the subgroup with the highest error as much as possible by optimizing, $\min \max P(\hat{Y} \neq Y | A = a)$

Notions of Fairness

Outside Computer Vision

- Adding additional Fairness measures to loss
- Enforcing protected attribute independent representation
- Data augmentation strategies

In Computer Vision

Biasing

Due to Sampling Inequalities

Mitigation Approaches

- Increasing Data Diversity
- Compensating Distribution Gaps with Synthetic Images
- Adaptive Resampling Methods

On Accuracy-based Fairness in Low and High-capacity Classifiers

Fairness Measure Satisfied

Any accuracy-based fairness measure is trivially satisfied **by a classifier with zero error.**

No Zero Error

For typical Low Dimensional datasets with large label volatility, zero error do not occur in practice

Computer Vision Datasets and Models

Empirically Shatterable: Even with random relabeling, achieving zero error on the training set is possible. Accuracy-based fairness trivially satisfied

Bias-Variance Decomposition for classification

Theoretical framework

Error

- Irreducible Label Noise, **N**
- Fit of regressor on the dataset, Bias **B**
- Variance **V**, Generalization Error

Definitions

- $N(x) = E_{y|x}[L(y, y_*(x)]$, induced by label disagreement
- $B(x) = L(y_*(x), y_m(x))$, Systematic model imperfection
- $V(x) = E_{D_n}[L(y_{,m}(x), f(x)]$, Difference from main prediction

Total Error

$$err_x = c1(x)N(x) + B(x) + c2(x)V(x)$$

For some $c1(x), c2(x) \in R$

Expected Fairness Violations

For two groups, A and B

- Expected Fairness Violation, E_{fair}

$$E_{fair} = |E_{x \in A}[err_x] - E_{x \in B}[err_x]|$$

Therefore, from previous definitions,

$$E_{fair} = |N_A + B_A + V_A - (N_B + B_B + V_B)$$

For low-capacity classifiers

- Variances are strongly dominated by biases. i.e $N_G + B_G \gg V_G$
- Approximated fairness violation

$$E_{fair} \approx |N_A + B_A - N_B - B\_B$$

For high-capacity classifiers

- No Label Disagreement, $V(X)$ vanishes
- Trains to Convergence, $B(X)$ vanishes
- Fairness violation dominated by Generalization error

$$E_{fair} \approx |V_A - V_B|$$

Rethinking Fairness Measures

Standard Fairness

Can be satisfied with random or constant classifiers

Performance Reduction

Reduces performance across all groups

Methods

Injecting Noise, Data Augmentation, Heuristics

Effective Evaluation

- Require improvements for disadvantaged groups
- Suggested metrics include accuracy/TPR of worse performing group (min-max fairness)

Improving accuracy on disadvantaged groups with synthetic data

Data Diversity to Improve Variance

Challenges & Solutions

Challenge	Solution
Decide which group requires augmentation	Deploys adaptive sampling strategies using held-out data
Generate High Fidelity In-Distribution Data	Use invertible GANs and latent space traversals to edit images
Reliably Augment and automatically label	g-SMOTE, a generalized synthetic minority oversampling technique, produces labeled images using GANs

Adaptive Sampling

Algorithm 1 Adaptive Sampling

Inputs:
Hyper-parameter $\lambda \in [0, 1]$
Train dataset $D_{\text{Train}} = \{(x_0, y_0), (x_1, y_1), \ldots\}, x_i \in \mathcal{X}, y_i \in \mathcal{Y}$
Evaluation dataset $D_{\text{Eval}} = \{(x_0^e, y_0^e), (x_1^e, y_1^e), \ldots\}, x_i^e \in \mathcal{X}, y_i^e \in \mathcal{Y}$
Classifier $c_\phi : \mathcal{X} \to \mathcal{Y}$ (parameterized by ϕ)
Initialize: $D_{\text{Aug}} := D_{\text{Train}}$
for $i = 1, \ldots, n_{\text{training steps}}$ **do**
  With probability λ, uniformly sample $(x_i, y_i) \in D_{\text{Train}}$, otherwise sample $(x_i, y_i) \in D_{\text{Aug}}$
  Update ϕ according to learning objective
  Determine weakest group based on learning objective and D_{Eval} and augment corresponding $x_{\text{Aug}}, y_{\text{Aug}}$ from that group
  $D_{\text{Aug}} \leftarrow D_{\text{Aug}} \cup \{(x_{\text{Aug}}, y_{\text{Aug}})\}$
end for

- Complex than just balancing group sizes
- Also accounts for group characteristics that can influence generalization performance.

Generalized SMOTE: g-SMOTE

Background

- **Task**: Generate synthetic images and attribute labels from the original dataset
- **SMOTE**: Synthetic Minority Over-sampling Technique
- **GAN**: Generative Adversarial Networks, allow images to be 'embedded' into their latent spaces

g-SMOTE

- SMOTE in GAN latent space
- Extends SMOTE to uniform sampling within a k-dimensional simplex formed by k of the m nearest neighbors, aimed at improving data diversity.
- Datapoint, its m-nearest neighbors with the same attribute chosen
- Latent points uniformly sampled from this simplex

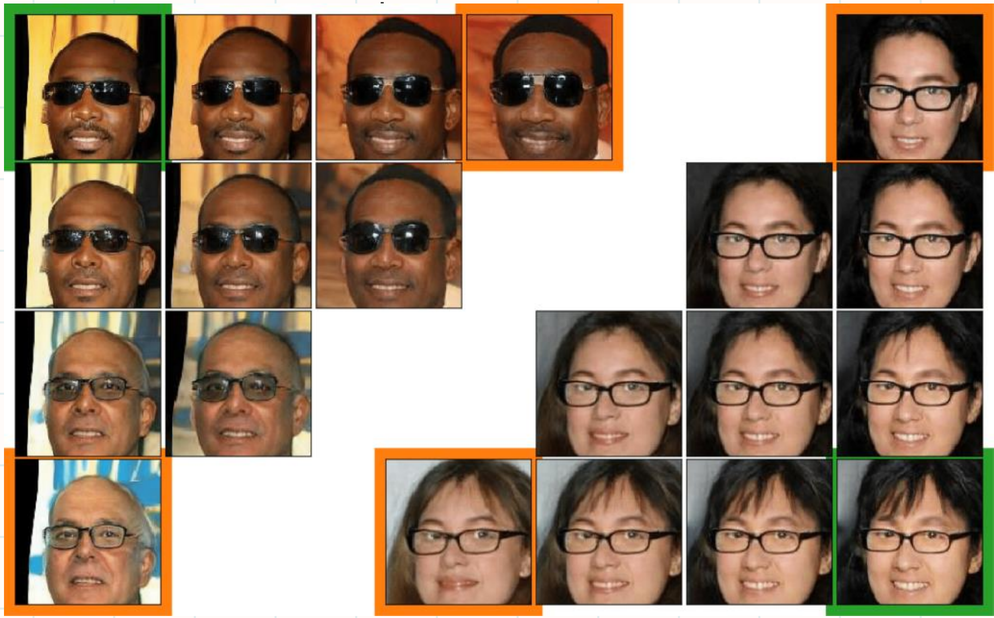
Assumption: Simplex covers a label-consistent volume in latent space

Generalized SMOTE: g-SMOTE

Background

- **Task**: Generate synthetic images and attribute labels from the original dataset
- **SMOTE**: Synthetic Minority Over-sampling Technique
- **GAN**: Generative Adversarial Networks, allow images to be 'embedded' into their latent spaces

g-SMOTE



Assumption: Simplex covers a label-consistent volume in latent space

Experiments

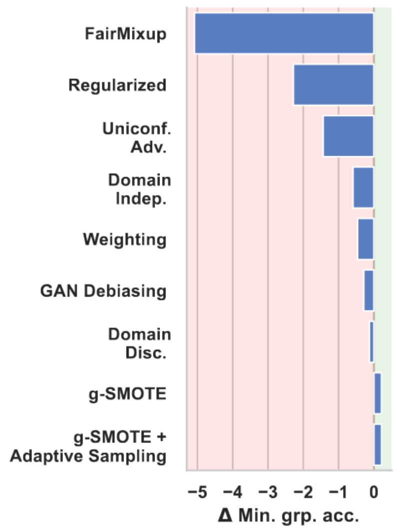
Configuration

- Base Model : ResNet-50
- Dataset : CelebA
- Selected Attribute Classification
- Trained with Adam and Rand Augment
- For GAN, InvGAN was used

Methods Compared

- Oversampling
- Domain Discriminative Training
- Domain Independent Models
- Adversarial Approaches
- Regularization
- GAN Based Debiasing

Key Finding



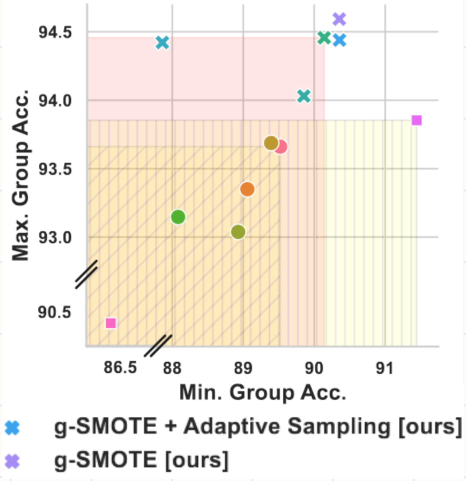
The only methods to increase the performance on the less accurate groups are g-SMOTE **with and without adaptive sampling**

Other Findings

- **Adaptive debiasing with g-SMOTE improves min-max performance**

  GAN allows for effective unsupervised data augmentation

- **Adaptive debiasing with g-SMOTE works with cross-sectional groups of multiple protected attributes**
- **g-SMOTE produces better data diversity than popular augmentation strategies**



Shaded Rectangles show Pareto Inefficient Regions

	No Augment	Rand Crop	Rand Rot.	Rand Flip	RandAugment
Without g-SMOTE	89.15	89.56	89.66	89.78	90.17
With g-SMOTE	**89.63**	**89.85**	**89.75**	**89.86**	**90.33**

Min Group Accuracy

Conclusion

Recommendations

- Evaluate a model using the **error of the worst performing group**
- Gather **more data** for the worst performing groups

Conclusion

- Fairness on unseen data is primarily a problem of **generalization**
- **Limitation**: Analysis only holds for accuracy-based fairness notions
- **Future Directions**: Hyperparameter Optimization, Neural Architecture Search, Data Augmentation

Wichita State University

Presented By:
Sreeraj Ramachandran

Mitigating Face Recognition Bias via Group Adaptive Classifier

Authors: Sixue Gong, Xiaoming Liu, Anil K. Jain

Introduction

Face Recognition Bias

FR bias is the uneven recognition performance w.r.t. demographic groups

Existing Mitigation Methods

Data or Loss Reweighting, Adaptive Clustering, Margin Loss Based Methods

Utilizing two types of Features

General Pattern: Shared by all faces, Differential Pattern: relevant to demographic attributes. On skewed datasets general pattern is convenient and leads to bias.

Introduction

Unbiased FR Model

- Should Rely on unique patterns for recognition of different groups
- Should Rely on General patterns of all faces for improved generalizability
- Proposed model, therefore, contains an adaptive model and loss

Adaptive Neural Networks

Adaptive Architectures

- Neural-Selection Hidden Layers
- Automatic CNN expansion

Dynamic Kernels

- Content Adaptive Convolutions
- Shape-Driven Kernels
- Automatic Receptive Fields

Attention Mechanisms

- Cross-Attention, Cross-Channel Communications

Methodology

Overview

Adaptive Layer

- Features maps convolved with unique kernels per group
- Followed by multiplication with adaptive attention maps

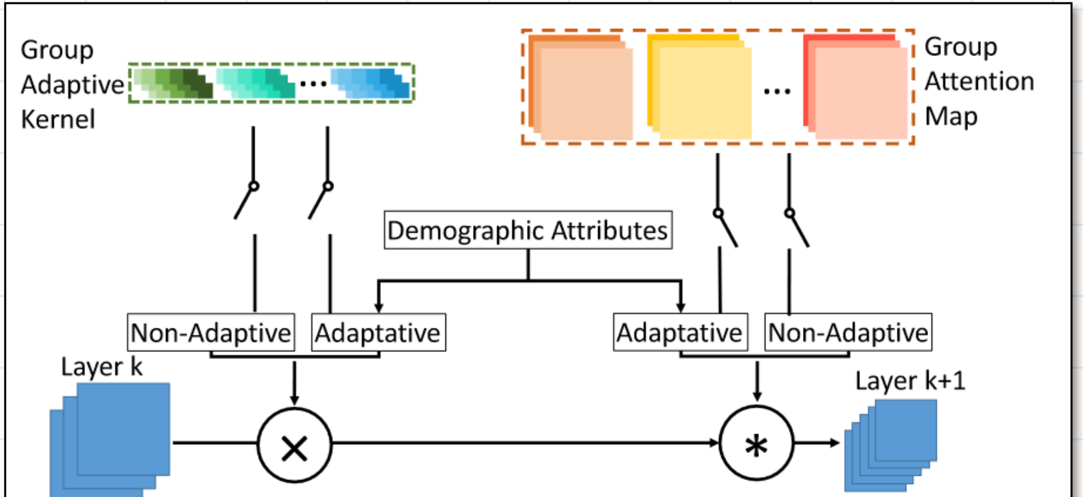
Automation Module

- Determines which layers adaptive kernels and attention should be applied
- Combined they obtain, demographic-differential features

Group Adaptive Classifer

Methodology

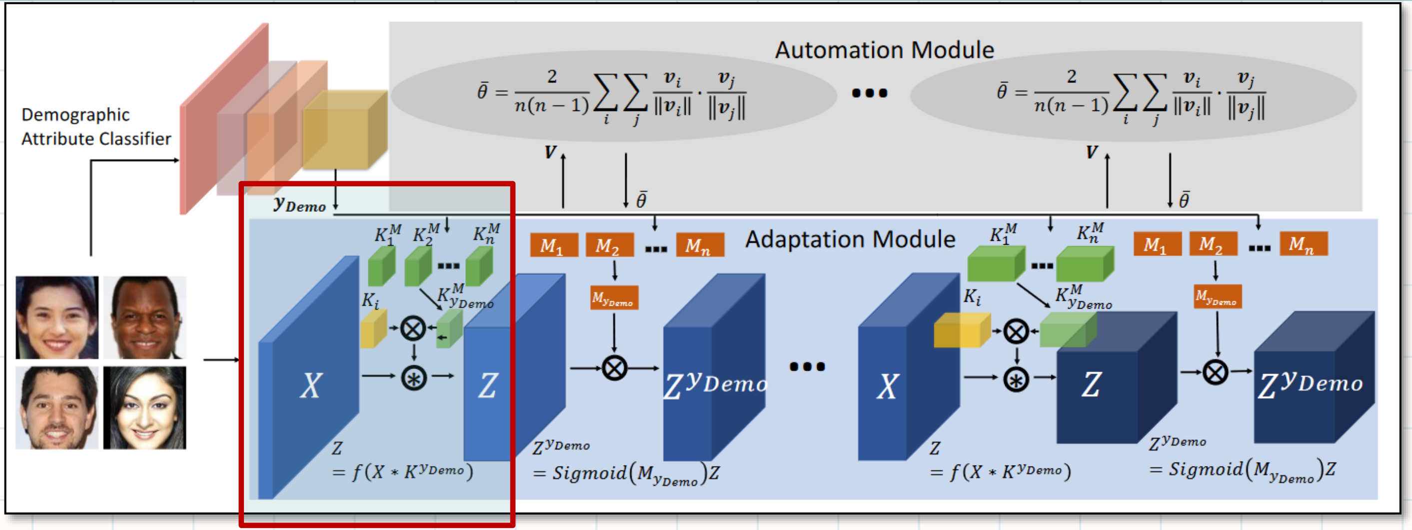
Overview



(c) GAC

Group Adaptive Classifer

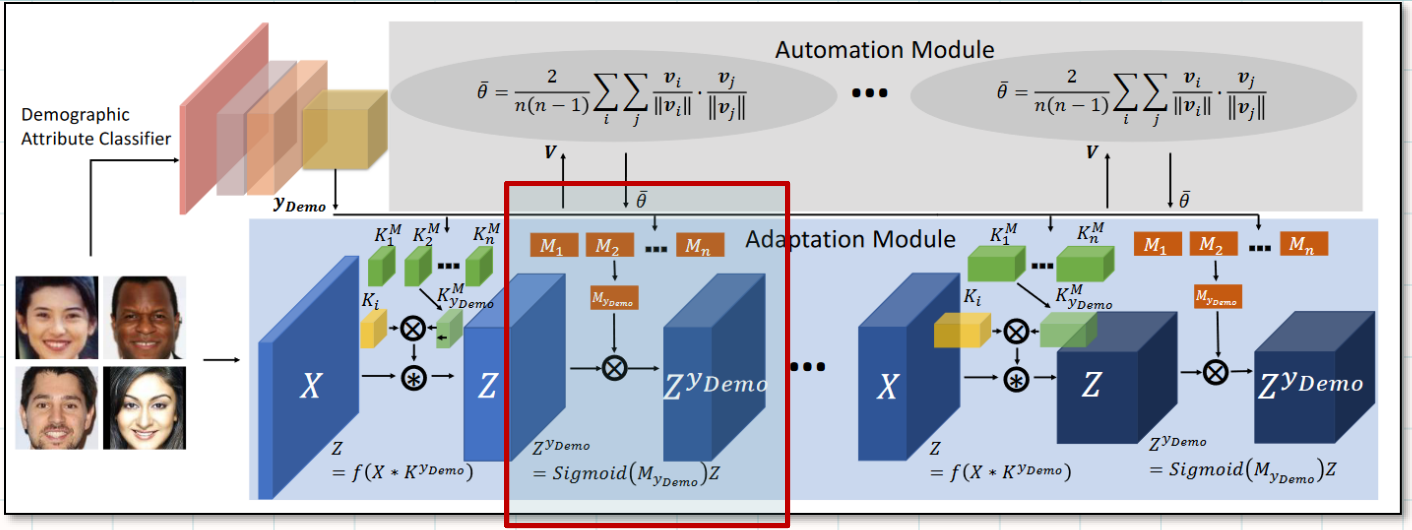
Adaptive units in GAC are constructed by demographic information and are automatically applied to corresponding layers



Adaptive Convolution

- **Standard Convolution** : Input → $X \in R^{c\times h^X \times w^X}$, convolved with single kernel, $K \in R^{k\times c\times h^K \times w^K}$
- Shares Kernel → Agnostic to demographic, results in limited capacity per group
- Introduce a trainable matrix of kernel masks, $K^M \in R^{n\times c\times h^X \times w^X}$, n → no. of groups
- Let y_{demo} be demographic label, then i^{th} channel of adaptive filter for group y_{demo}

$$K_i^{y_{demo}} = K_i \otimes K_{y_{demo}}^M$$

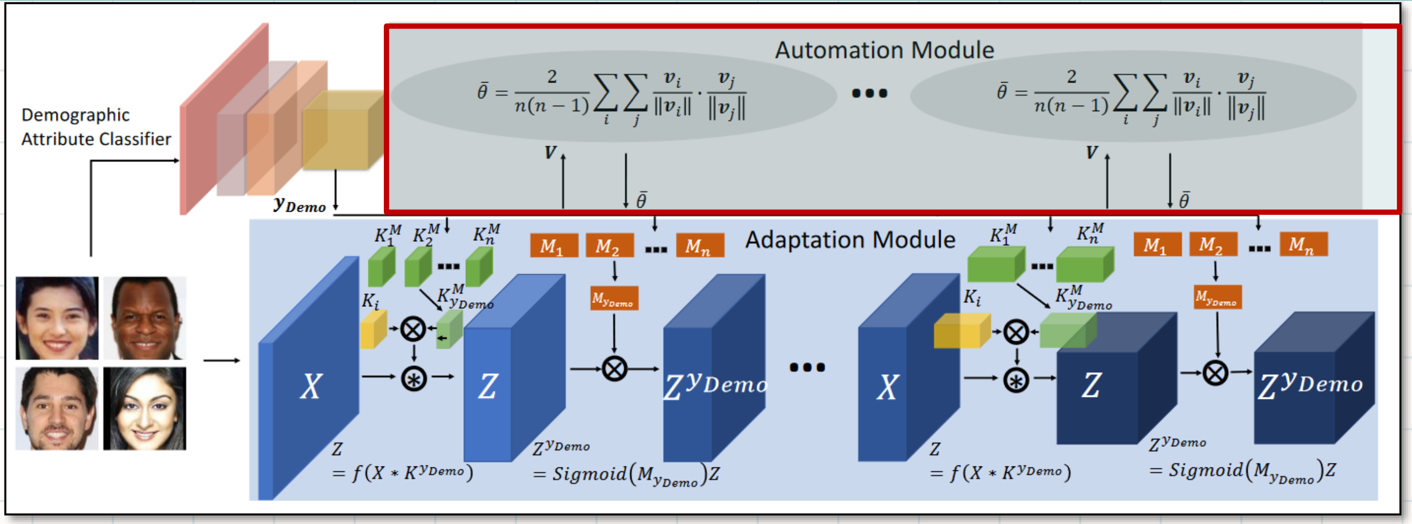


Adaptive Attention

- In Adaptive Convolution, kernel mask broadcasted along channels ➔ Weight selection spatially varied but channel-wise joint.
- Introduces **Channel-wise Attention Maps**, $M \in R^{n\times k}$
- Given y_{demo} and feature map Z, i^{th} channel of feature map is given by

$$Z_i^{y_{demo}} = sigmoid\left(M_{y_{demo}^i}\right)Z_i$$

WICHITA STATE UNIVERSITY



Automation Module

- Adding an adaptation module to every layer is inefficient
- The kernel masks from the adaptation module are used to calculate the average pairwise similarity score.
- Based on a predefined threshold τ, merge n kernels groupwise
- When τ decreases, more layers will be adaptive

De-biasing Objective Function

- Regress loss function to narrow the gap of the intra-class distance between demographic groups
- Let $\boldsymbol{r}_{ijg} = g(I_{ijg}, \boldsymbol{w})$, be the feature representation of I_{ijg}, i^{th} image of subject j in group g
- Average intra-class distance of subject j

$$Dist_{jg} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{r}_{ijg} - \boldsymbol{\mu}_{jg})^T(\boldsymbol{r}_{ijg} - \boldsymbol{\mu}_{jg})$$

This allows us to lower the difference of intra-class distance by

$$L_{bias} = \frac{\lambda}{Q \times n}\sum_{g=1}^{n}\sum_{j=1}^{Q}\left|Dist_{jg} - \frac{1}{n}\sum_{g=1}^{n} Dist_g\right|$$

$Dist_g$ → intra-class distance for all subjects in group g
λ → coefficient for the de-biasing objective
Q → number of total subjects in group g

Experiments

Configuration

- Datasets : RFW, BUPT-Balanced Face
- 50-layer ArcFace Architecture
- Classification Loss : CosFace
- Trained Gender Classifier combining 5 other datasets (ResNet-18)

Performance Metrics

- Demographic Parity is improper
- Used Standard Deviation of Performance across groups
- Biasness → error between the average and the performance on group
- Average Accuracy

Key Finding

Method	White	Black	East Asian	South Asian	Avg	STD
RL-RBN	96.27	95	94.82	94.68	95.19	0.63
ACNN	96.12	94	93.67	94.55	94.58	0.94
PFE	**96.38**	**95.17**	94.27	94.6	95.11	0.93
ArcFace	96.18	94.67	93.72	93.98	94.64	0.96
CosFace	95.12	93.93	92.98	92.93	93.74	0.89
DebFace	95.95	93.67	94.33	94.78	94.68	0.83
GAC	96.2	94.77	**94.87**	**94.98**	**95.21**	**0.58**

Performance comparison with SOTA on the RFW protocol

Ablation Studies - Adaptive Strategies

Adaptive mechanisms, Number of Convolutional layers, and Demographic Information

Method	White	Black	East Asian	South Asian	Avg	STD
Baseline	96.18	93.98	93.72	94.67	94.64	1.11
GAC-Channel	95.95	93.67	94.33	94.78	94.68	0.83
GAC-Kernel	96.23	94.4	94.27	94.8	94.93	0.78
GAC-Spatial	95.97	93.2	93.67	93.93	94.19	1.06
GAC-CS	96.22	93.95	94.32	**95.12**	94.65	0.87
GAC-CSK	96.18	93.58	94.28	94.83	94.72	0.95
GAC-(τ=0)	96.18	93.97	93.88	94.77	94.7	0.92
GAC-(τ=-0.1)	**96.25**	94.25	94.83	94.72	95.01	0.75
GAC-(τ=-0.2)	96.2	**94.77**	**94.87**	94.98	**95.21**	**0.58**

Observations

- Baseline Model Most Biased
- Spatial Attention mitigates at the cost of accuracy
- Combining Adaptive kernels with attention increases parameter count, lowering performance
- Small τ may increase redundant adaptive layers, while large τ may result in lack of capacity

Ablation Studies - Depths and Demographic Labels

Demographic labels

- Ground-truth from dataset
- Estimated using pretrained model
- Randomly Assigned

Observations

- Successfully reduces STD at various depths
- Noise and Bias in labels impair performance
- Biasness : Random > Estimated > Ground Truth

Method	White	Black	East Asian	South Asian	Avg	STD
Number of Layers						
ArcFace-34	96.13	93.15	92.85	93.03	93.78	1.36
GAC-ArcFace-34	96.02	94.12	94.1	94.22	94.62	0.81
ArcFace-50	96.18	93.98	93.72	94.67	94.64	1.11
GAC-ArcFace-50	96.2	94.77	94.87	94.98	95.21	0.58
ArcFace-100	96.23	93.83	94.27	94.8	94.78	0.91
GAC-ArcFace-100	96.43	94.53	94.9	95.03	95.22	0.72
Race/Ethnicity Labels						
Ground-truth	96.2	94.77	94.87	94.98	95.21	0.58
Estimated	96.27	94.4	94.32	94.77	94.94	0.79
Random	95.95	93.1	94.18	94.82	94.5	1.03

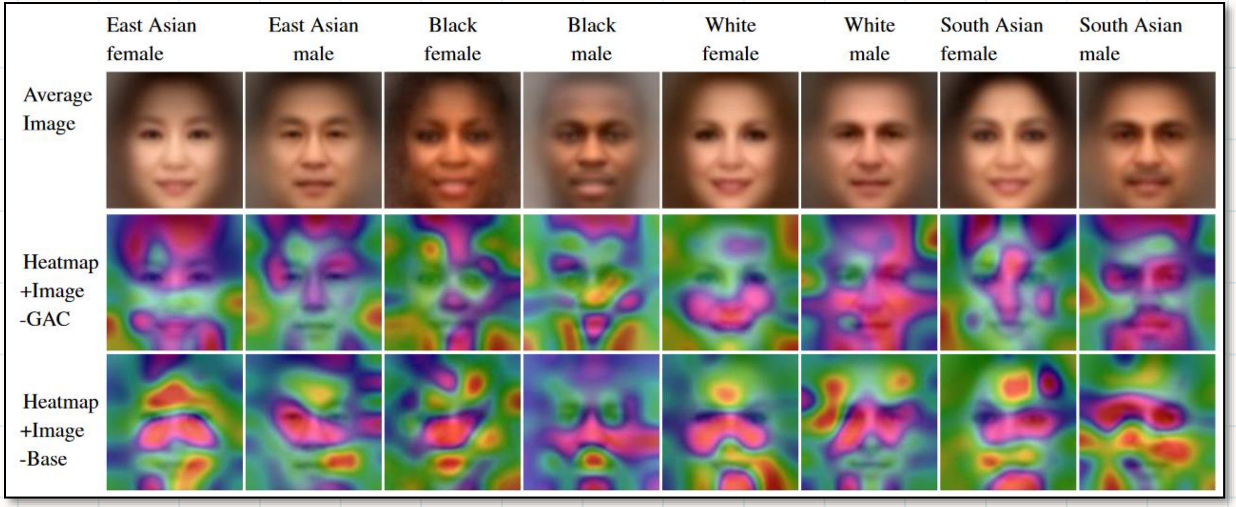
Method	Gender	White	Black	East Asian	South Asian	Avg	STD
Baseline	Male	97.49 ±0.08	96.94 ± 0.26	97.29 ± 0.09	97.03 ± 0.13	96.96 ± 0.03	0.69 ± 0.04
	Female	97.19 ± 0.10	97.93 ± 0.11	95.71 ± 0.11	96.01 ± 0.08		
AL+Manual	Male	98.57 ± 0.10	98.05 ± 0.17	98.50 ± 0.12	98.36 ± 0.02	98.09 ± 0.05	0.66 ± 0.07
	Female	98.12 ± 0.18	**98.97 ± 0.13**	96.83 ± 0.19	**97.33 ± 0.13**		
GAC	Male	**98.75 ± 0.04**	**98.18 ± 0.20**	**98.55 ± 0.07**	98.31 ± 0.12	**98.19 ± 0.06**	**0.56 ± 0.05**
	Female	**98.26 ± 0.16**	98.80 ± 0.15	**97.09 ± 0.12**	**97.56 ± 0.10**		

Verification Accuracy (%) of 5-fold cross-validation on 8 groups of RFW

Effectiveness of Automation Module

- AL+Manual adds adaptive kernels and attention maps to a subset of layers
  - First block in residual unit is AdaptiveConv and Attention applied on output from last block
- Automatic adaptation is more effective in enhancing the discriminability and fairness of face representations

Visualization and Analysis on Bias of FR



Gradient Weighted Class Activation Maps from 43rd convolutional layer of GAC and Baseline

- Salient regions of GAC demonstrate more diversity on faces from different groups
- The higher diversity of heatmaps in GAC shows the variability of parameters in GAC across groups.

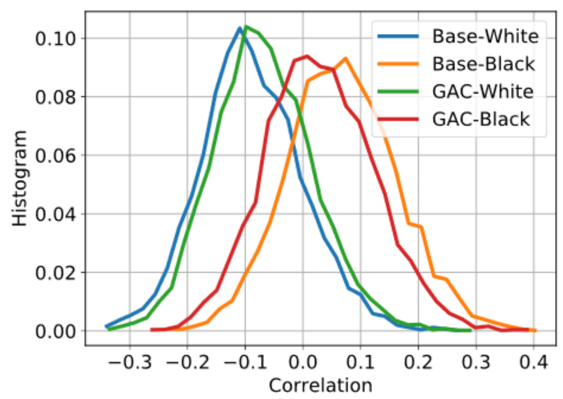
Visualization and Analysis on Bias of FR

Effectiveness of Automation Module

- **Assumption** : Statistics of neighbors of a given point(representation) reflects certain properties of its manifold(local geometry)

- Base-White representations show **lower inter-class correlation** than Base-Black → White group are over-represented by the baseline

- GAC-White and GAC-Black shows **more similarity** in their correlation histograms



Pair-wise correlation of face representations in same race

For local geometry, it's the ratio of minimum inter-subject distance to maximum intra-subject distance is computed

- GAC's racial ratio distributions **align closely with the reference**, indicating **less bias**.

Conclusion

- The paper addresses demographic bias in face recognition via a fair face representation.
- It introduces a **Group Adaptive Classifier (GAC)** to enhance demographic group representation robustness.
- GAC incorporates **adaptive convolution kernels** and **channel-wise attention maps**.
- An **automation module** is included to decide when to use adaptations.
- Results show that demographic-specific adaptive layers **improve face representation**, balancing performance across all groups.

References

- TODO

Thanks!