ENHANCING FAIRNESS OF VISUAL ATTRIBUTE PREDICTORS

Tobias Hänel[1], Nishant Kumar[1], Dmitrij Schlesinger[1], Mengze Li[2], Erdem Ünal, Abouzar Eslami[2], Stefan Gumhold[1]

Chair for Computer Graphics and Visualization, TU Dresden, Germany[1]; Carl Zeiss Meditec AG, Munich, Germany[2]

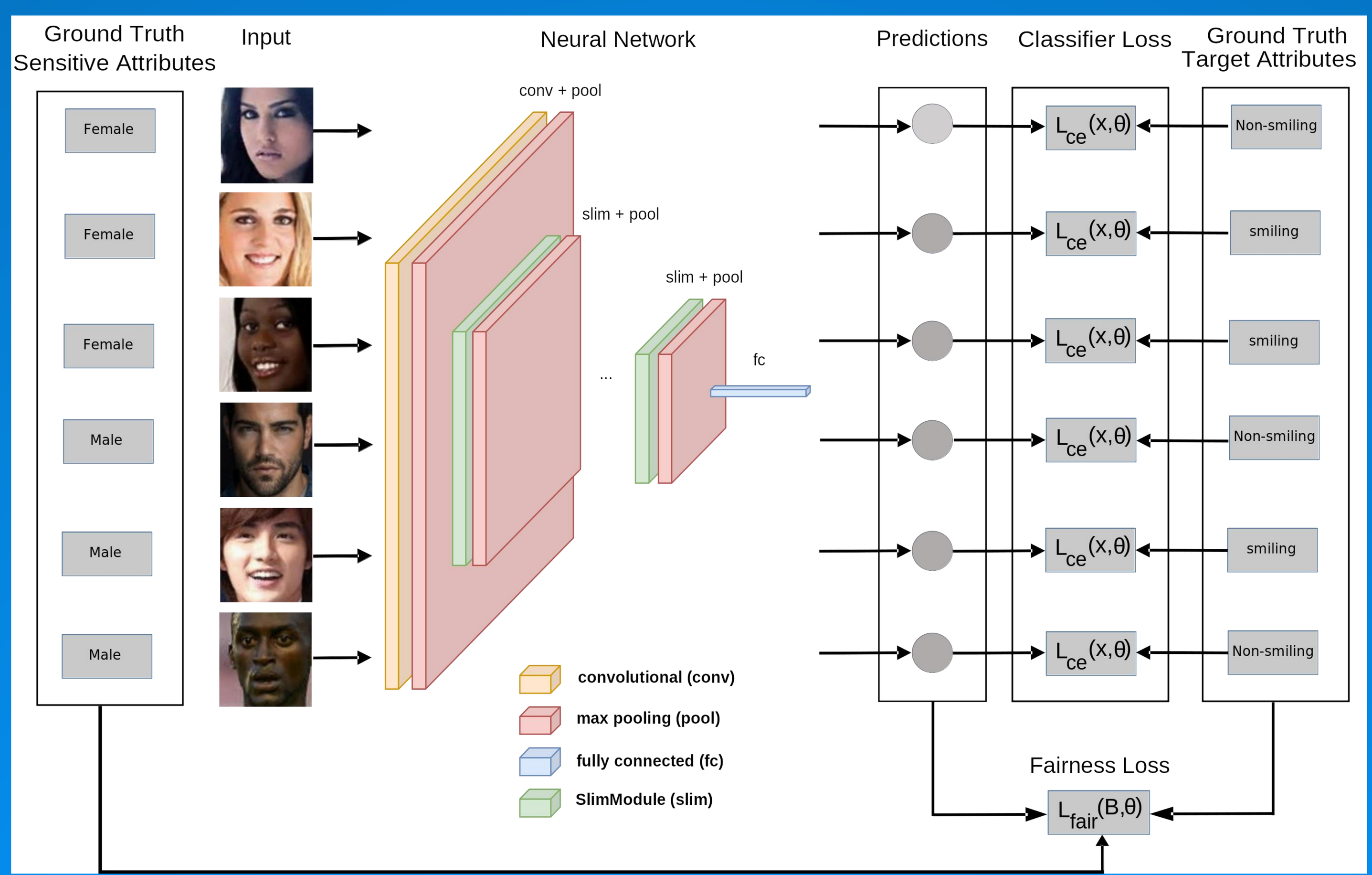
I. Motivation

- *Problem*: Bias is present in our society (e.g. credit limits for women, criminal justice for PoC)
- *Reason*: Human decisions are influenced by existing prejudices
- *Observation*: Recent machine learning (ML) algorithms can aid impartial decision making (e.g. unbiased recruitment automation)
- *New Problem*: ML algorithms are prone to biases - dependency on data quality
- *Idea*:
  - Achieve *algorithmic fairness* with existing biased data sets
  - Learn fairness during training to reduce bias w.r.t. *sensitive attributes* (e.g. age, gender, ethnicity)

II. Contributions

- Implementation of *Demographic Parity* (DP) and *Equalized Odds* (EO) fairness notations as differentiable loss functions for categorical variables
- Development of a novel performance based *Intersection-over-Union* (IoU) loss
- Verifying experiments on publicly available data sets:
  - Facial attribute prediction on *CelebA*
  - Age group estimation on *UTKFace*
  - Disease classification on *SIIM-ISIC Melanoma*

III. Training System



Proposed fairness aware training system

- *Training data* $T = (x_1, x_2, \dots, x_{|T|})$ consisting of $|T|$ images $x \in \mathcal{X}$
- *Ground-truth sensitive attribute labels* $y_s^*(x) \in \{1 \dots K_s\}$ (e.g. male or female)
- *Ground-truth target attribute labels* $y_t^*(x) \in \{1 \dots K_t\}$ (e.g. smiling or non-smiling)
- *Predicted target attribute labels* $y_t(x) \in \{1 \dots K_t\}$ (e.g. smiling or non-smiling)
- *Trainable classifier* $p_\theta(y_t|x)$ conditional probability distribution
  - *Learnable parameter* θ (e.g. CNN network weights)
- *Loss function* $L(\theta)$
  - *Cross-entropy loss* L_{ce}
  - *Weighting coefficient* λ
  - *Fairness loss* L_{fair}

$$L(\theta) = \mathbb{E}_{B \subset T}\left[\sum_{x \in B} L_{ce}(x,\theta) + \lambda \cdot L_{fair}(B,\theta)\right]$$

- *Image batches* $B \subset T$ → Fairness estimation and mini-batch gradient descent

IV. a) Fairness Losses – Demographic Parity

- *Requirements*: Predictions shouldn't depend on sensitive attribute ($y_t \perp y_s^*$)

$$p(y_t = a | y_s^* = b) = p(y_t = a) \ \forall a \in \{1 \dots K_t\}, b \in \{1 \dots K_s\}$$

- $L_{dp}^{\ell 2}$ *loss*: Sum of squared probability differences

$$L_{dp}^{\ell 2}(\theta) = \sum_{a,b} [p_\theta(y_t = a | y_s^* = b) - p_\theta(y_t = a)]^2$$

- L_{dp}^{mi} *loss*: $D_{KL}\big(p_\theta(y_t, y_s^*) \parallel p_\theta(y_t) \cdot p_\theta(y_s^*)\big)$ = Mutual information (MI) between target attribute predictions and sensitive attribute ground-truth $I(y_t; y_s^*)$

$$L_{dp}^{mi}(\theta) = \sum_{a,b} p_\theta(y_t = a, y_s^* = b) \cdot \log \frac{p_\theta(y_t = a, y_s^* = b)}{p_\theta(y_t = a) \cdot p_\theta(y_s^* = b)} = H(y_t) + H(y_s^*) - H(y_t, y_s^*)$$

IV. b) Fairness Losses – Equalized Odds

- *Requirements*: Predictions shouldn't depend on sensitive attribute for a fixed value of the ground-truth target attribute ($(y_t \perp y_s^*)|y_t^*$)

$$p(y_t = a | y_t^* = b, y_s^* = c) = p(y_t = a | y_t^* = b) \ \forall a, b \in \{1 \dots K_t\}, c \in \{1 \dots K_s\}$$

- $L_{eo}^{\ell 2}$ *loss*: Sum of squared probability differences

$$L_{eo}^{\ell 2}(\theta) = \sum_{a,b,c} [p_\theta(y_t = a | y_t^* = b, y_s^* = c) - p_\theta(y_t = a | y_t^* = b)]^2$$

- L_{eo}^{mi} *loss*: Sum of MI scores between target attribute predictions and sensitive attribute ground-truth conditioned on ground truth target attribute labels

$$L_{eo}^{mi}(\theta) = \sum_{a} [H(y_t | y_t^* = a) + H(y_s^* | y_t^* = a) - H(y_t, y_s^* | y_t^* = a)]$$

IV. c) Fairness Losses – Intersection over Union

- *Goal*: Similar prediction performances for each sensitive attribute class
- *Performance measure*: Ratio of correct predictions to all occurrences of a target attribute label (predictions or ground truth)

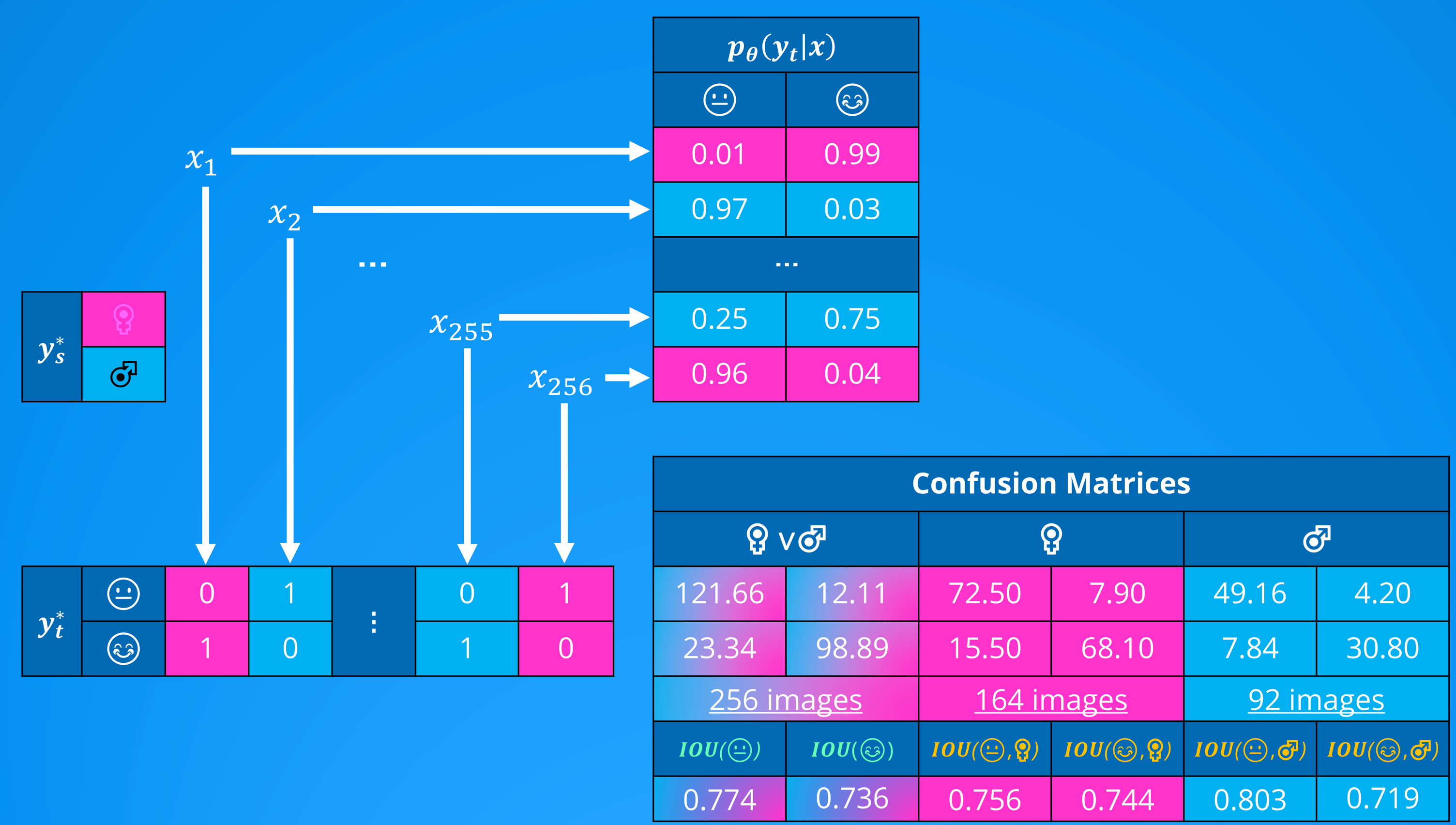
$$IOU_\theta(a) = \frac{p_\theta(y_t = a \land y_t^* = a)}{p_\theta(y_t = a \lor y_t^* = a)}$$

- *Conditioning*: Consider only samples of a specific sensitive attribute class

$$IOU_\theta(a,b) = \frac{p_\theta((y_t = a \land y_t^* = a) \land y_s^* = b)}{p_\theta((y_t = a \lor y_t^* = a) \land y_s^* = b)}$$

- *Reduction*: Average over target attribute labels → overall and sensitive IOUs
- L_{IOU} *loss*: Sum of squared differences between overall and sensitive IOUs

$$L_{iou}(\theta) = \sum_b \left[\left(\frac{1}{K_t}\sum_a IOU_\theta(a,b)\right) - \left(\frac{1}{K_t}\sum_a IOU_\theta(a)\right)\right]^2$$



IOU computation for a batch of 256 images

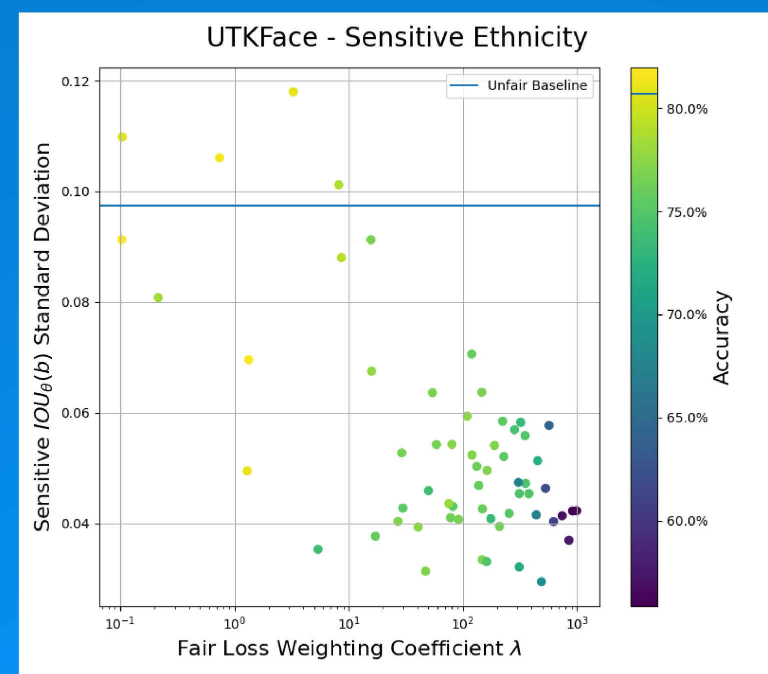
V. Experimental Results

CelebA – Sensitive Male						
Loss	Accuracy	L_{iou}	$L_{eo}^{\ell 2}$	L_{eo}^{mi}	$L_{dp}^{\ell 2}$	L_{dp}^{mi}
L_{ce}	0.902	8.73e-4	4.89e-3	5.12e-3	1.77e-2	8.46e-3
L_{iou}	**0.903**	7.32e-5	8.59e-4	4.26e-4	2.51e-3	1.20e-3
$L_{eo}^{\ell 2}$	0.902	**1.35e-5**	**1.78e-4**	**7.71e-5**	**1.36e-4**	**6.45e-5**
L_{eo}^{mi}	0.899	2.37e-5	2.24e-4	1.03e-4	8.40e-4	4.00e-4
$L_{dp}^{\ell 2}$	0.899	4.28e-5	3.75e-3	1.87e-3	1.57e-4	7.43e-5
L_{dp}^{mi}	0.901	5.28e-5	7.73e-3	3.96e-3	7.15e-4	3.40e-4

Results for facial attribute prediction with the fairness losses on CelebA

- *CelebA data set*:
  - >200K celebrity images
  - 40 binary attributes (e.g. Wearing Hat, Smiling)
- *Network*: SlimCNN (memory-efficient CNN)
- *Attributes*: y_t^* = Smiling and y_s^* = Male
- *Training*:
  - *Baseline model*: L_{ce} for 100 epochs
  - *Fair models*: Baseline → $L_{ce} + \lambda \cdot L_{fair}$ for 25 epochs

VI. Hyperparameter Optimization



Results for the HPO of the weighting coefficient λ

- *Motivation*: Investigate relationship between weighting coefficient λ, prediction performance and fairness
- *HPO Objective*: Fairness ≙ standard deviation of sensitive IOU scores for different sensitive attribute labels

$$\sigma_{IOU}(\lambda) = \sqrt{\frac{1}{K_s - 1}\sum_{i=1}^{K_s}\left(\left(\frac{1}{K_t}\sum_j^{K_t} IOU_\theta(a_i, b_j)\right) - \left(\frac{1}{K_s K_t}\sum_j^{K_t} IOU_\theta(a_i, b_k)\right)\right)^2}$$

- *UTKFace data set*:
  - >20k facial images
  - 3 attributes (age, gender and ethnicity)
- *Network*: EfficientNet (scalable CNN)
- *Attributes*: y_t^* = Age Group and y_s^* = Ethnicity

TECHNISCHE UNIVERSITÄT DRESDEN

software campus

ACCV 2022 Macau

ZEISS

Computergraphik und Visualisierung

github.com/nish03/FVAP (MIT License)