MultiFair: Model Fairness With Multiple Sensitive Attributes (IEEE 2025)

Sehyun Park

April 30, 2025

Seoul National University - IDEA Lab.

Introduction

- Previous mixup-based fairness methods assume only one binary sensitive attribute.
- **MultiFair** extends mixup to fuse data across all sub-groups, creating a "neutral domain" that removes information about multiple sensitive attributes at once.
- In this paper, authors design three mixup schemes that balance information fusion across attributes while retaining distinct visual features critical for training valid models.

Mixup operation for single attribute

- **Mixup operation:**
  - $\mathcal{Y} = \{0, 1\}, \mathcal{S} = \{0, 1\}$: label and sensitive-attribute spaces.
  - $\mathcal{X}^{train} = \bigcup_{y \in \{0,1\}, s \in \{0,1\}} \mathcal{X}^{train}_{y,s}$: the entire training dataset, where $\mathcal{X}^{train}_{y,s}$ denotes the subset with label y and sensitive attribute s.
  - For each $x_i \in \mathcal{X}^{train}$ with label y_i and sensitive attribute s_i, randomly choose a counterpart x_i^c from $\mathcal{X}^{train}_{y_i,\, 1-s_i}$.
  - Then, the mixed sample x_i^m is composed as follows:

$$x_i^m := \lambda x_i + (1 - \lambda) x_i^c \quad , \forall i = 1, \ldots, n$$

where λ is a weighting parameter randomly set during training.

Mixup operations for multiple attributes

- Smiling classification using the CelebA dataset.
- Sensitive attribute : {Male, Young, Eyeglasses}
- Only smiling images $(y = 1)$ are considered.
- $\mathcal{X}^{train}$:

(1,0,0) (1,1,1) (1,0,1) (0,1,0) (0,1,1) (1,1,0)

$\Rightarrow x_i$:

(1,0,1)

1) Mixup in Turn

1: At each training step, select one sensitive attribute.
=> {Young}

2: Randomly select a counterpart based on {Young}.
=> x_i^c :

3: Mixup
=> x_i^m:

2) Mixup in distance

1: Randomly select a counterpart from those with the largest difference in the sensitive attribute vector.
=> x_i^c :

2: Mixup
=> x_i^m:

3) Mixup via interpolations

1: For each sensitive attribute, randomly select an image with a different attribute value.
=> {Male} {Young} {Eyeglasses}

2: Generate a counterpart with interpolations
=> x_i^c :

3: Mixup
=> x_i^m:

Mixup operations for multiple attributes

1) Mixup in Turn

- $\mathcal{S}^j = \{0, 1\}$: j-th sensitive attribute, for $j = 1, \ldots, K$.
- At each training step, only one attribute $\mathcal{S}^j$ is selected.
- $\mathcal{X}^{train} = \bigcup_{y \in \{0,1\}, s \in \{0,1\}} \mathcal{X}^{train}_{y,s,j}$: the full training dataset, where $\mathcal{X}^{train}_{y,s,j}$ denotes the subset with label y and j-th sensitive attribute s.
- For each $x_i \in \mathcal{X}^{train}$ with label y_i and j-th sensitive attribute s_{ij}, randomly select a counterpart as

$$x_i^c \in \mathcal{X}^{train}_{y_i, 1-s_{ij}, j}.$$

- Then, the mixed sample x_i^m is composed as follows:

$$x_i^m := \lambda x_i + (1 - \lambda) x_i^c \quad , \forall i = 1, \ldots, n$$

where λ is a weighting parameter randomly set during training.

Mixup operations for multiple attributes

2) Mixup in Distance

- $V(x_i)$: the mapping function that outputs all K sensitive attributes of x_i.
  e.g. $V(x_i) = [0, 1, 1, \cdots, 0] \in \{0, 1\}^K$
- For $x_i \in \mathcal{X}^{train}$, randomly select a counterpart as

$$x_i^c \in \text{argmax}_{x' \in \mathcal{X}^{train}} \left| V(x_i) - V(x') \right|.$$

- Then, the mixed sample x_i^m is composed as follows:

$$x_i^m := \lambda x_i + (1 - \lambda) x_i^c \quad , \forall i = 1, \ldots, n$$

  where λ is a weighting parameter randomly set during training.

Mixup operations for multiple attributes

3) Mixup via Interpolations

- $V_j(x_i)$: the mapping function that outputs j-th sensitive attribute of x_i.
- For $x_i \in \mathcal{X}^{train}$, randomly select K samples as

$$x_{i,j}^c \in \mathcal{X}_{y,1-V_j(x_i),j}^{train}, \quad \forall j = 1, \ldots, K.$$

- Then generate counterpart of x_i as

$$x_i^c = \sum_{j=1}^{K} \frac{1}{K} x_{i,j}^c.$$

- Then, the mixed sample x_i^m is composed as follows:

$$x_i^m := \lambda x_i + (1-\lambda) x_i^c \quad , \forall i = 1, \ldots, n$$

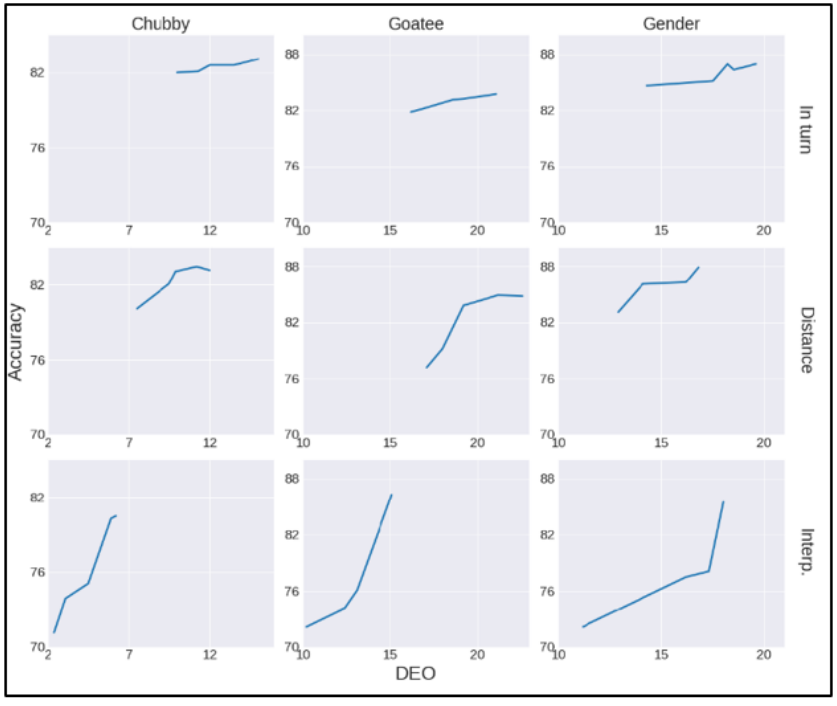
where λ is a weighting parameter randomly set during training.

Experiment Result

〈Results for CelebA dataset〉

Considered attributes	Methods	Mean Acc↑	Δ_{DP}↓	Δ_{EO}↓	s_0 Acc↑	s_1 Acc↑
The presence of chubby	Biased	75.5	8.2	14.6	71.4	79.6
	Resample	74.3	8.4	12.4	72.2	76.4
	EOP	72.1	4.4	10.7	72	73.7
	Adv	78.7	3	11.8	78.6	78.8
	FCRO	79.2	2.9	7.6	79.4	79.0
	Random	80	9.2	12.8	75.4	84.6
	Distance	**83.1**	6.6	9.9	79.9	86.4
	In turn	82	6.5	10	78.7	85.2.4
	Interp.	80.3	**2.7**	**5.9**	78.9	80.3
The presence of goatee	Biased	77.6	10.4	26	72.4	82.8
	Resample	79.1	7.3	15.8	75.3	82.9
	EOP	80.1	12.9	17.5	78.6	81.6
	Adv	81.8	7	18	79.4	84.2
	FCRO	81.6	6.1	**13.2**	79.6	83.6
	Random	80.8	10	18	75.8	85.8
	Distance	83.8	9.8	19.2	78.9	88.8
	In turn	83.1	10.3	18.6	77.8	88.2
	Interp.	**86.3**	**6.1**	15.1	80.2	86.3
Gender	Biased	80.1	5.4	28.6	77.4	82.8
	Resample	79.3	4.7	23.1	78.5	80.1
	EOP	77.3	5.7	17.2	76.8	77.8
	Adv	82.6	3	21.6	82.2	83
	FCRO	83.6	2.5	19.3	83.0	84.2
	Random	85.7	5.8	20.6	82.8	88.6
	Distance	**87.7**	3.9	**16.2**	85.7	89.6
	In turn	86.9	5.8	18.2	84.1	89.8
	Interp.	85.5	**2.4**	18	85.8	85.1

〈Fairness-accuracy trade-off across different attribute〉



End

Experimental setup

- Datasets
  - CelebA
    - 202,599 images of real celebrity faces, all centrally aligned to a frontal view.
    - Each image comes with 40 binary attributes (e.g., eyeglasses, smile, gender) and 5 facial landmarks (both eyes, nose tip, and mouth corners).
    - This paper conducts training using information from three sensitive attributes. (*gender*, *chubby*, and *goatee*)
- Baseline Methods
  - Preprocessing
    - Resampling method
  - In-Processing
    - Adversarial training methods (Adv)
    - Fair classification orthogonal representation (FCRO)
  - Preprocessing
    - EOs postprocessing methods (EOP)