Self-Supervised Fair Representation Learning without Demographics

Junyi Chai
chai28@purdue.edu

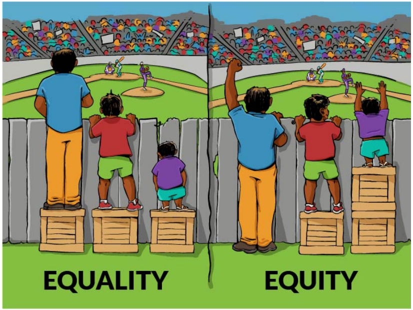
Xiaoqian Wang
joywang@purdue.edu

November 30, 2022

PURDUE UNIVERSITY
Elmore Family School of Electrical and Computer Engineering

Introduction

As machine learning systems are increasingly used for automated decision making with social impact, discrimination across different demographic groups has become an important concern.

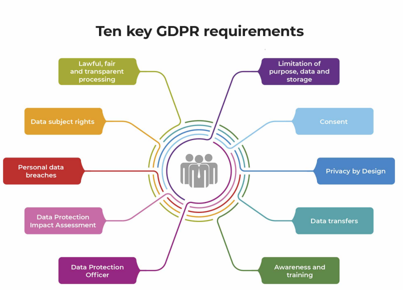


Elmore Family School of Electrical and Computer Engineering

Introduction

However, in real-world scenarios, due to privacy or legal concern, it might be infeasible to collect or use the sensitive information.

Under such scenarios, conventional methods on fairness would fail to work.



PURDUE UNIVERSITY
Elmore Family School of Electrical and Computer Engineering

Introduction

Much of current literature on fairness without demographics focuses on fully supervised setting.

Instead, we consider a more general extension: fairness without demographics and with partially available labels.

Our goal: contrastive learning method with gradient-based reweighing to learn fair representations without demographics.

PURDUE UNIVERSITY
Elmore Family School of Electrical and Computer Engineering

Method

Contrastive learning:

$$\mathcal{L}_{ctr}(\tilde{\boldsymbol{x}}_i;\theta) = -\log \frac{\exp(\mathrm{sim}(f_\theta(\tilde{\boldsymbol{x}}_i), f_\theta(\tilde{\boldsymbol{x}}_i^{\mathrm{pos}}))/\tau)}{\sum_{j\neq i} \exp\left(\mathrm{sim}\left(f_\theta(\tilde{\boldsymbol{x}}_i), f_\theta(\tilde{\boldsymbol{x}}_j)\right)/\tau\right)}.$$

Max-Min fairness:

$$l(k,\theta) = \left[\frac{1}{k}\sum_{i=1}^{2N}\left[\mathcal{L}_{ctr}(\tilde{x}_i;\theta) - \lambda(k,\theta)\right]_+ + \lambda(k,\theta)\right].$$

Problem: false negative pairs during sampling

PURDUE UNIVERSITY
Elmore Family School of Electrical and Computer Engineering

Instead, we consider to minimize the top-k validation loss:

$$
\begin{aligned}
& l^{\mathrm{val}}(k, \theta, \omega) \\
&= \left[\frac{1}{k} \sum_{j=1}^{M} \left[\mathcal{L}_{cls}\left(g_\omega(f_\theta(\boldsymbol{x}_j)), \boldsymbol{y}_j \right) - \lambda^{\mathrm{val}}(k, \theta, \omega) \right]_+ + \lambda^{\mathrm{val}}(k, \theta, \omega) \right].
\end{aligned}
$$

$$
\begin{aligned}
\theta^*(v) &= \arg\min_\theta \frac{1}{2N} \left[\sum_{i=1}^{2N} v_i \mathcal{L}_{ctr}(\tilde{\boldsymbol{x}}_i; \theta) \right], \\
v^*, \omega^* &= \arg\min_{v \geq 0, \omega} l^{\mathrm{val}}(k, \theta^*(v), \omega).
\end{aligned}
$$

PURDUE UNIVERSITY
Elmore Family School of Electrical and Computer Engineering

Weight approximation

Estimation via cosine similarity:

$$u_{t,i} = \left(\nabla_\theta l_t^{\mathrm{val}}\right)^\top \nabla_\theta l_{t,i}.$$

Intra-batch normalization:

$$\hat{v}_{t,i} = \max\left(u_{t,i}, 0\right),$$
$$v_{t,i} = \frac{2n\hat{v}_{t,i}}{\sum_{i'=1}^{2n} \hat{v}_{t,i'} + \delta\left(\sum_{i'=1}^{2n} \hat{v}_{t,i'}\right)}.$$

PURDUE UNIVERSITY
Elmore Family School of Electrical and Computer Engineering

Theoretical analysis

Assumption

We have the following two assumptions.

1. *The partial derivative of validation loss* l^{val} *with respect to* θ *is Lipschitz continuous with constant* L*, i.e.,* $\nabla^2_{\omega\theta} l^{val}$ *and* $\nabla^2_{\theta\theta} l^{val}$ *are upper-bounded by* L*.*
2. *The contrastive loss* l *has* σ*-bounded gradients w.r.t.* θ*.*

PURDUE UNIVERSITY
Elmore Family School of Electrical and Computer Engineering

Theoretical analysis

Theorem

Under Assumption 1, at iteration t*, let the learning rate of contrastive encoder* f *satisfies* $\alpha_{1,t} \leqslant \frac{4\sigma^2 L \sum_i \beta_{t,i}^2}{n\sum_i \left(\beta_{t,i}^2 - 2\gamma_{t,i}\beta_{t,i}\right)}$*, and the learning rate of linear classifier satisfies* $\alpha_{2,t} \leq \min\left(\frac{2}{L}, \frac{\sum_i \beta_{t,i}^2}{L\sum_i \gamma_{t,i}\beta_{t,i}}\right)$*, where*

$$\gamma_{t,i} = \|\nabla_\omega l_t^{val}\| \|\nabla_\theta l_{t,i}\|, \quad \beta_{t,i} = \left(\left(\nabla_\theta l_{t,i}\right)^\top \nabla_\theta l_t^{val}\right),$$

then the validation loss will monotonically decrease until convergence.

Experiments

Table 5: Results on the CelebA dataset with gender as sensitive attribute and attractive as label.

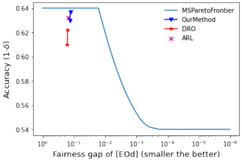
	Methods	Accuracy (%)	Disparate Impact (%)	Equalized Odds (%)
Methods with Correct Demographics	Postprocessing (gender)	78.32±0.87	11.24±1.88	8.67±2.34
	TAC (gender)	79.32±0.61	13.21±1.67	10.23±2.96
Methods with Wrong Demographics	Postprocessing (age)	77.43±1.83	14.01±2.56	18.42±1.60
	TAC (age)	78.82±0.71	17.31±2.68	19.63±2.23
Methods without Demographics	Fully supervised baseline	**80.43±1.62**	18.62±3.29	22.37±5.82
	Contrastive learning baseline	79.13±0.57	18.21±4.03	20.64±5.45
	DRO	76.38±2.66	15.33±3.09	17.61±4.43
	ARL	76.43±1.37	14.44±2.19	16.83±2.76
	Our method	77.63±0.79	**14.32±1.89**	**16.17±1.97**

Table 6: Results on the CelebA dataset with age as sensitive attribute and gender as label.

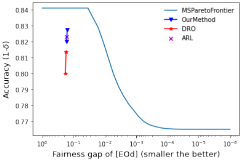
	Methods	Accuracy (%)	Disparate Impact (%)	Equalized Odds (%)
Methods with Correct Demographics	Postprocessing (age)	86.83±0.86	11.17±1.59	8.13±3.03
	TAC (age)	88.12±0.92	9.45±2.09	5.27±2.48
Methods with Wrong Demographics	Postprocessing (smiling)	86.32±0.72	14.01±1.28	12.67±2.15
	TAC (smiling)	87.76±0.96	14.33±2.93	12.25±1.75
Methods without Demographics	Fully supervised baseline	**89.74±0.84**	16.75±4.85	14.44±4.80
	Contrastive learning baseline	87.43±0.84	16.25±2.53	14.43±4.93
	DRO	72.43±2.63	15.21±1.73	13.44±2.34
	ARL	85.54±0.73	14.67±3.59	12.59±1.34
	Our method	86.93±0.72	**11.34±2.50**	**10.82±2.37**

Experiments

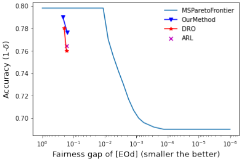
Fairness-accuracy trade-off:



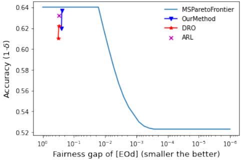
(a) COMPAS (gender)



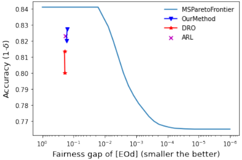
(b) Adult (gender)



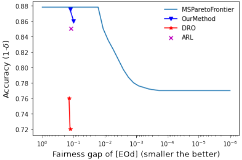
(c) CelebA (gender)



(d) COMPAS (race)



(e) Adult (race)



(f) CelebA (age)

Figure: Pareto frontier on Adult, CelebA and COMPAS dataset.

PURDUE UNIVERSITY
Elmore Family School of Electrical and Computer Engineering

Summary

Semi-supervised fair representation learning without demographics

Top-k average loss as surrogate fairness constraint

Gradient similarity based weight assignment

Convergence guarantee

PURDUE UNIVERSITY
Elmore Family School of Electrical and Computer Engineering

Thank you

Elmore Family School of Electrical and Computer Engineering