Transferring Fairness under Distribution Shifts via Fair Consistency Regularization

Jinwon Park

January 30, 2023

Seoul National University

Table of Contents

Introduction

- Most of fair algorithms are based on a assumption of an identical training and test distribution.
- However, such an assumptions are often violated in real-world applications
- In this paper, the author proposes a method to transfer model fairness under distribution shift with a fair consistency regularization as the key component.

Table of Contents

Notation

- X, A, Y: random variables of input features, sensitive attributes and label
- $\mathcal{X}, \mathcal{A}, \mathcal{Y}$: sample space of input features, sensitive attributes and label
- $Y^1, Y^2, \ldots, Y^K$: underlying factors ($A, Y \in \{Y^1, Y^2, \ldots, Y^K\}$)
  $\rightarrow$ underlying factors other than A and Y are nuisance factors
- $g : \mathcal{X} \rightarrow \mathcal{Y}$: classifier
- g_{tc}: teacher classifier
- S_a^y, T_a^y: sample space of $X|_{A=a,Y=y}$ in source domain and target domain
  ($S = \cup_y \cup_a S_a^y$, $T = \cup_y \cup_a T_a^y$, $a \in \{0, 1\}$, and $y \in \{0, 1\}$)
- U_a^y, U: group distribution and distribution on the entire data
  (i.e. group is a set of data that has same sensitive attribute and label)

Fairness Metric

- equalized odds

$$\Delta_{odds} = \frac{1}{2}\sum_{y=0}^{1}\left|P(\hat{Y}=y|A=0,Y=y) - P(\hat{Y}=y|A=1,Y=y)\right|$$

- variance of group accuracy

$$V_{acc} = Var(\{P(\hat{Y}=y|A=a,Y=y), \forall a, y\})$$

Table of Contents

Underlying data generation process

- (Assumption 1) Underlying data generation process
  - assume that the data is generated from a latent generative model as;

$$y^{1:K} \sim P(Y^{1:K}) \text{ and } x \sim P(X|Y^{1:K} = y^{1:K})$$

  - the generative model is fixed;

$$P_S(X|Y^{1:K} = y^{1:K}) = P_T(X|Y^{1:K} = y^{1:K})$$

  - but the marginal distribution of factors varies in two domains;

$$P_S(Y^{1:K}) \neq P_T(Y^{1:K})$$

  - hence, result in distribution shifts;

$$P_S(Y^{1:K}, X) \neq P_T(Y^{1:K}, X)$$

Domain Shift

- (Definition 1) Domain Shift
  If at least one nuisance factor Y^i has different sample space in two domains, $\exists y^i \in \mathcal{Y}_T^i$, but $y^i \notin \mathcal{Y}_S^i$, results in $P_S(Y^{1:K}) \neq P_T(Y^{1:K})$ and $P_S(Y^{1:K}, X) \neq P_T(Y^{1:K}, X)$

- Under domain shift, the source model has never seen the data with factor values that only exist in the target domain

Table of Contents

A Sufficient Condition for Transferring Fairness

- (Assumption 2) Separability of the input

$$S_a^y \cup S_{a\prime}^y = T_a^y \cup T_{a\prime}^y = S_a^y \cup T_{a\prime}^y = \emptyset,\ \forall y, a \neq a\prime$$

$$S_a^y \cup S_{a\prime}^{y\prime} = T_a^y \cup T_{a\prime}^{y\prime} = S_a^y \cup T_{a\prime}^{y\prime} = \emptyset,\ \forall a, a\prime, y \neq y\prime$$

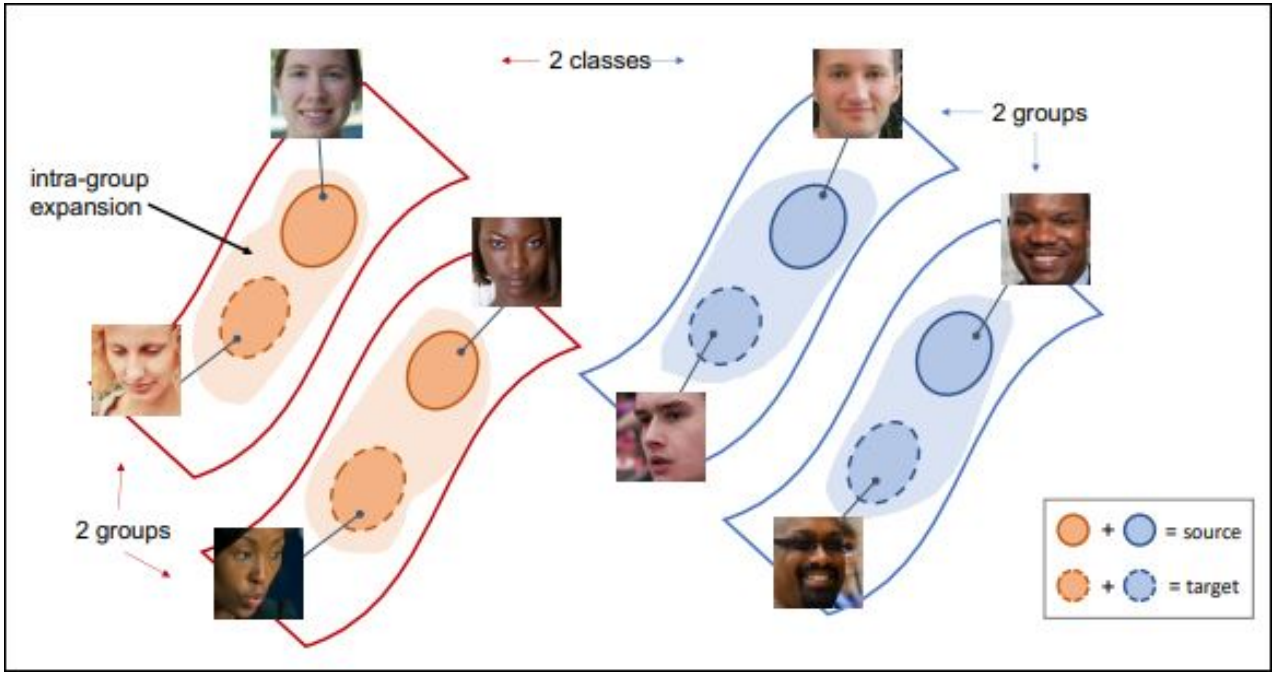
(groups are separated by label and sensitive attribute)

- (Definition 2) Neighbor
  $\rightarrow$ Let $\mathcal{T}$ denote a set of input transformations and define the transformation set of x as $\mathcal{B}(x) \triangleq \{x\prime | \exists t \in \mathcal{T}, \text{s.t.} \|x\prime - t(x)\| \leq r\}$.
  For any $x \in S_a^y \cup T_a^y$, define the neighbor of x as $\mathcal{N}(x) := (S_a^y \cup T_a^y) \cap \{x\prime | \mathcal{B}(x) \cap \mathcal{B}(x\prime) \neq \emptyset\}$ and define the neighbor of a set $V \in \mathcal{X}$ as $\mathcal{N}(V) := \cup_{x \in V \cap (\cup_y \cup_a S_a^y \cup T_a^y)} \mathcal{N}(x)$.
  $\rightarrow$ Only consider neighbors that have the same class and sensitive attribute (i.e., from the same group).

A Sufficient Condition for Transferring Fairness

- (Assumption 3) Intra-group expansion
  U_a^y satisfies (α, c)-multiplicative expansion for some constant $\alpha \in (0, 1)$ and $c > 1$, if for all $V \subset U_a^y$ with $P_{U_a^y}(V) \leq \alpha$, the following holds;

$$P_{U_a^y}(\mathcal{N}(V)) \geq \min\{cP_{U_a^y}(V), 1\}$$



A Sufficient Condition for Transferring Fairness

- (Theorem 1) Guarantee fairness
  Suppose we have a teacher classifier g_{tc} with bounded unfairness such that $|\varepsilon_{U_a^y}(g_{tc}) - \varepsilon_{U_{a\prime}^{y\prime}}(g_{tc})| \leq \gamma, \forall a, a\prime \in \mathcal{A}$ and $y, y\prime \in \mathcal{Y}$. We assume intra-group expansion where U_a^y satisfies $(\bar{\alpha}, \bar{c})$-multiplicative expansion and $\varepsilon_{U_a^y}(g_{tc}) \leq \bar{\alpha} < 1/3$ and $\bar{c} > 3, \forall a, y$. We define $c \triangleq \min\{1/\bar{\alpha}, \bar{c}\}$, and set $\mu \leq \varepsilon_{U_a^y}(g_{tc}), \forall a, y$. If we train our classifier with algorithm

$$\min_{g \in G} \max_{a,y} R_{U_a^y}(g), \qquad \text{s.t. } L_{U_a^y}(g, g_{tc}) \leq \mu \quad \forall a, y$$

  then, the error and unfairness of the optimal solution $\hat{g}$ on the distribution U are bounded with

$$\varepsilon(\hat{g}) \leq \frac{2}{c-1}\varepsilon_U(g_{tc}) + \frac{2c}{c-1}R_U(\hat{g})$$

$$\Delta_{odds} \leq \frac{2}{c-1}(\gamma + \mu + c \max_{a,y} R_{U_a^y}(\hat{g}))$$

Fair Consistency Regularization

- Two challenges in realizing theorem 1 (Guarantee fairness)
  1. we need a high quality teacher model, but the model is trained only with labeled source data
     $\rightarrow$ leveraging self-training paradigm that updates the teacher model

  2. existing consistency regularization do not take fairness into consideration
     $\rightarrow$ propose a fair consistency regularization

Fair Consistency Regularization

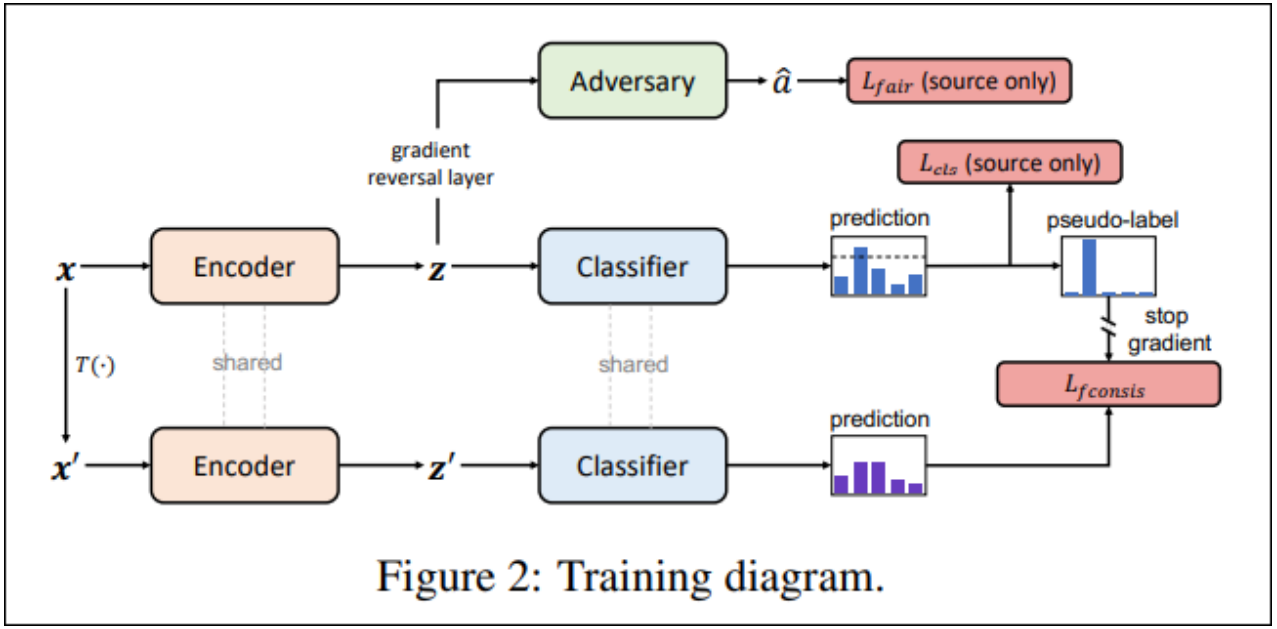


Figure 2: Training diagram.

- $L_{fair} = \sum_{(a,y)\in\{0,1\}^2} \frac{1}{|D_a^y|} \sum_{x\in D_a^y} |h(f(x)) - a|$
- $L_{fconsis}(g) = \sum_{y=0}^{1} \sum_{a=0}^{1} \lambda_a^y L_a^y(g)$

  $L_a^y(g) = \frac{1}{\sum_{x_a^y}} \sum_{x_a^y} 1(\max(g_{tc}(x_a^y)) \geq \tau) H(\text{argmax}(g_{tc}(x_a^y)), g(t(x_a^y)))$

  $\hat{\lambda}_a^y = \frac{1}{\sum_{x_a^y} 1(\max(g_{tc}(x_a^y)) \geq \tau)}, \quad \lambda_a^y = \frac{\hat{\lambda}_a^y}{\sum_{a,y} \hat{\lambda}_a^y}$

Experiments

Method	Source			Target		
	Acc	Unfairness		Acc	Unfairness	
		V_{acc}	Δ_{odds}		V_{acc}	Δ_{odds}
Base	92.85±0.49	2.30±0.97	4.81±0.69	74.49±0.83	5.79±3.49	9.90±1.27
Laftr	93.24±0.41	1.19±0.46	2.44±0.51	74.35±1.46	6.92±0.72	9.79±1.54
CFair	92.51±0.22	1.76±0.53	4.75±0.85	73.53±0.89	7.51±0.73	7.26±1.95
Laftr+DANN	91.33±0.08	2.12±1.72	2.70±0.67	74.28±1.63	6.25±2.59	8.27±2.11
CFair+DANN	90.89±0.76	2.01±0.70	4.43±1.36	74.62±1.06	6.23±0.90	5.26±2.07
Laftr+FixMatch	96.62±0.06	0.77±0.21	2.23±0.44	83.87±0.48	8.21±0.67	9.32±1.01
CFair+FixMatch	96.13±0.53	1.28±0.53	2.78±0.74	83.11±0.49	7.87±1.86	7.89±0.40
Ours (w/ Laftr)	96.08±0.07	0.96±0.39	2.59±0.35	85.52±0.40	2.82±0.87	5.70±0.52
Ours (w/ CFair)	95.65±0.22	1.56±0.37	3.85±0.97	84.48±0.42	2.88±0.99	5.43±0.65

Table 1: Transfer fairness and accuracy from UTKFace to FairFace

Q&A?