Learning Adversarially Fair and Transferable Representations

(2018)

David Madras, Elliot Creager, Toniann Pitassi, Richard Zemel

Summary

Contributions

For a given prediction tasks, we are involved in two steps: (1) acquiring data in a suitable form and specifying an algorithm that have a good accuracy on the acquired data. In the first step: a data owner needs to find the correct representation of the data, in the second step: a vendor wants to have the highest prediction scores without introducing any bias. Both parties are constrained by each other, the data owner must only collect data that will yield fair predictors and the vendor needs to construct predictor with relatively high utility, while maintaining fairness. THe authors propose to view this as an adversarial framework, where the data owner's choices are criticized by a critic if we have unfair solutions, and propose the appropriate fairness metrics to use in such a framework. The objective to have fair classifiers trained on the produced representations, and being able to *fair transfer* by admitting fair predictors on unseen tasks.

Method

Fairness. The objective in fairness is to have stable predictions $\hat{Y}$ that are not biased toward one group A or an other, formulated as follows $P(\hat{Y} = 1|A = 0) = P(\hat{Y} = 1|A = 1)$, called *demographic parity*. But in case there is a difference in the ground truth between the two groups: $P(Y = 1|A = 0) \neq P(Y = 1|A = 1)$, the fairness criterion can hold. A solution is to define *equalized odds* by also conditioning on the ground truth labels: $P(\hat{Y} \neq Y|A = 0, Y = y) = P(\hat{Y} \neq Y|A = 1, Y = y)$.

Adversarially Fair Representations.

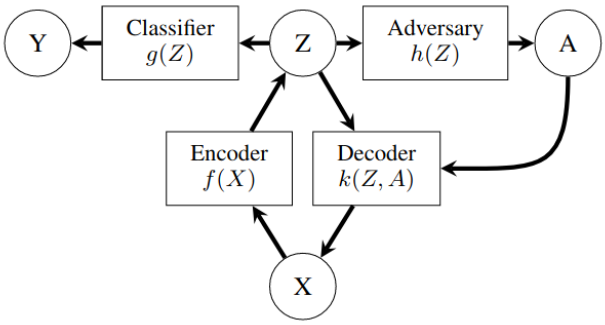


Figure 1. Model for learning adversarially fair representations. The variables are data X, latent representations Z, sensitive attributes A, and labels Y. The encoder f maps X (and possibly A - not shown) to Z, the decoder k reconstructs X from (Z, A), the classifier g predicts Y from Z, and the adversary h predicts A from Z (and possibly Y - not shown).

The objective is to generate representation Z using an encoder f, that can be reconstructed by a decoder k, used for classification by a classifier g, and able to high the sensitive attribute A form an adversary h. We have the following components:

- Data tuples (X, A, Y) in $\mathbb{R}^n$,$\{0,1\}$ and $\{0,1\}$.
- Encoder $f : \mathbb{R}^n \to \mathbb{R}^m$
- Classifier and adversary $g, h : \mathbb{R}^m \to \{0,1\}$
- Decoder $k : \mathbb{R}^m \times \{0,1\} \to \mathbb{R}^n$

With such a formulation, the encoder is what the data owner really wants; this yields the representations which will be sold to vendors. When the encoder is learned, the other two parts of the model ensure that the representations respond appropriately to each type of vendor: the classifier ensures utility by simulating an indifferent vendor (only wants good performances) with a prediction task, and the adversary ensures fairness by simulating an adversarial vendor with discriminatory goals. It is important to the data owner that the models adversary be as strong as possible, if it is too weak, the owner will underestimate the unfairness enacted by the adversarial vendor.

The training objective is to minimize the reconstruction loss L_{Dec} of the decoder, the classification loss L_C and maximize the discriminative loss L_{Adv} of the adversary: $\min_{f,g,k} \max_{h} \mathbb{E}_{X,Y,A}[L(f,g,h,k)]$, with the loss as follows:

$$L(f,g,h,k) = \alpha L_C(g(f(X,A)),Y) + \beta L_{Dec}(k(f(X,A),A),X) + \gamma L_{Adv}(h(f(X,A)),A)$$

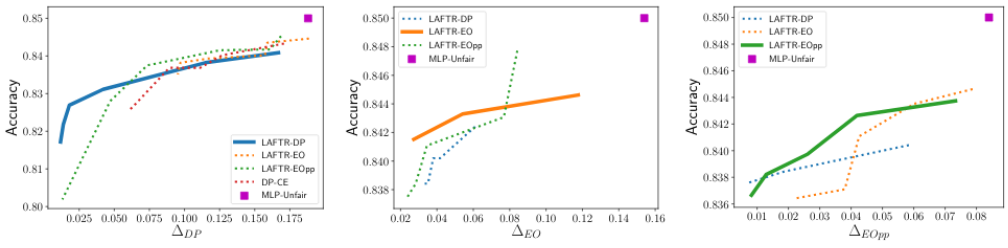
The networks are trained by alternating the gradient decent and ascent steps as in normal GaN training. The adversarial loss depend on either we're on *demographic parity* case, with two sensitive groups $\mathcal{D}_0$ and $\mathcal{D}_1$:

$$L_{Adv}^{DP}(h) = 1 - \sum_{i\in\{0,1\}} \frac{1}{|\mathcal{D}_i|} \sum_{(x,a)\in\mathcal{D}_i} |h(f(x,a)) - a|$$

Or for *equalized odds* on each sensitive group-label combination: $\mathcal{D}_0^0, \mathcal{D}_1^0, \mathcal{D}_0^1, \mathcal{D}_1^1$, where $\mathcal{D}_i^j = \{(x,y,a) \in \mathcal{D} | a = i, y = j\}$:

$$L_{Adv}^{EO}(h) = 2 - \sum_{(i,j)\in\{0,1\}^2} \frac{1}{\left|\mathcal{D}_i^j\right|} \sum_{(x,a)\in\mathcal{D}_i^j} |h(f(x,a)) - a|$$

Results



(a) Tradeoff between accuracy and Δ_{DP} (b) Tradeoff between accuracy and Δ_{EO} (c) Tradeoff between accuracy and Δ_{EOpp}

Figure 2. Accuracy-fairness tradeoffs for various fairness metrics (Δ_{DP}, Δ_{EO}, Δ_{EOpp}), and LAFTR adversarial objectives ($L_{Adv}^{DP}, L_{Adv}^{EO}, L_{Adv}^{EOpp}$) on fair classification of the Adult dataset. Upper-left corner (high accuracy, low Δ) is preferable. Figure 2a also compares to a cross-entropy adversarial objective (Edwards & Storkey, 2016), denoted DP-CE. Curves are generated by sweeping a range of fairness coefficients γ, taking the median across 7 runs per γ, and computing the Pareto front. In each plot, the bolded line is the one we expect to perform the best. Magenta square is a baseline MLP with no fairness constraints. see Algorithm 1 and Appendix B.

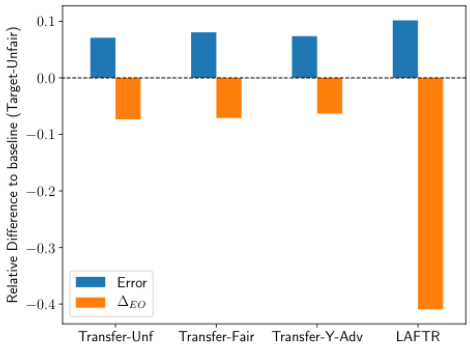


Figure 3. Fair transfer learning on Health dataset. Displaying aver age across 10 transfer tasks of relative difference in error and Δ_{EO} unfairness (the lower the better for both metrics), as compared to a baseline unfair model learned directly from the data. -0.10 means a 10% decrease. Transfer-Unf and -Fair are MLP's with and withou fairness restrictions respectively, Transfer-Y-Adv is an adversaria model with access to the classifier output rather than the under lying representations, and LAFTR is our model trained with the adversarial equalized odds objective.

Table 1. Results from Figure 3 broken out by task. Δ_{EO} for each of the 10 transfer tasks is shown, which entails identifying a primary condition code that refers to a particular medical condition. Most fair on each task is bolded. All model names are abbreviated from Figure 3; "TarUnf" is a baseline, unfair predictor learned directly from the target data without a fairness objective.

TRA. TASK	TARUNF	TRAUNF	TRAFAIR	TRAY-AF	LAFTR
MSC2A3	0.362	0.370	0.381	0.378	**0.281**
METAB3	0.510	0.579	**0.436**	0.478	0.439
ARTHSPIN	0.280	0.323	0.373	0.337	**0.188**
NEUMENT	0.419	0.419	0.332	0.450	**0.199**
RESPR4	0.181	0.160	0.223	0.091	**0.051**
MISCHRT	0.217	0.213	0.171	0.206	**0.095**
SKNAUT	0.324	**0.125**	0.205	0.315	0.155
GIBLEED	0.189	0.176	0.141	0.187	**0.110**
INFEC4	0.106	0.042	0.026	**0.012**	0.044
TRAUMA	0.020	0.028	0.032	0.032	**0.019**